ABSTRACT

Title of dissertation: Data Fusion based on the Density Ratio Model

Chen Wang
Doctor of Philosophy, 2018

Dissertation directed by: Professor Benjamin Kedem
Department of Mathematics

A vast amount of the statistical literature deals with a single sample coming from a distribution where the problem is to make inferences about the distribution by estimation and testing procedures. Data fusion is a process of integrating multiple data sources in the hope of getting more accurate inference than that provided by a single data sources, the expectation being that fused data are more informative than the individual original inputs. This requires appropriate statistical methods which can provide inference by using multiple data sources as input. The Density Ratio Model is a model which allows semiparametric inference about probability distributions from fused data. In this dissertation, we will discuss three different types of problems based on the Density Ratio Model. We will discuss the situation where there is a system of sensors, each producing data according to some probability distribution. The parametric connection between the distributions allows various hypothesis tests including that of equidistribution, which are very helpful in detecting abnormalities in mechanical systems. Another example of a data fusion problem is the small area estimation where borrowing strength occurs by using all

data from all areas where information is available. Real data can be fused with other real data, or even with artificial data. Thus, a given sample can be fused with computer-generated data giving rise to the concept of out of sample fusion(OSF). We will see that this approach is very helpful when estimating a small threshold exceedance probability when the sample size is not large enough and consisting of values below the threshold.

This dissertation is organized as follows: In Chapter One, an overview of the Density Ratio Model will be given. Chapter Two discusses applications of the data fusion idea in mechanical quality control. Chapter Three discusses the small area estimation problems where we propose a new way to estimate small area quantiles. Chapter Four gives an overview of Extreme Value Theory. Chapter Five describes the ideas of Out of Sample Fusion (OSF) and Repeated Out of Sample Fusion (ROSF). Chapter Six gives a new iteration method to estimate small threshold exceedance probability.

Data Fusion based on the Density Ratio Model

by

Chen Wang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Benjamin Kedem, Chair/Advisor
Professor Paul J. Smith
Professor Yanir A. Rubinstein
Professor Xin He
Professor Frank B. Alt

Dedication

To my family

Acknowledgments

I owe my sincere gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to extend my sincere gratitude to my adviser, Professor Benjamin Kedem for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past few years. I am also deeply grateful for his instructive advice and useful suggestions regarding my thesis. It has been a great pleasure to work with and learn from such an extraordinary individual.

I would also like to thank Professor Paul J. Smith, Professor Xin He, Professor Frank B. Alt and Professor Yanir A. Rubinstein for agreeing to serve on my thesis committee and for sparing their invaluable time reviewing the manuscript. Their expertise and insight are very valuable and greatly enriched my work.

I would also like to thank all professors that have helped me in my graduate studies. Professor Eric V. Slud taught me mathematical statistics and statistical computing, which helped me in understanding many fundamental questions in statistics and how to applying theory to practice by using R. Professor Bill Rand and Professor Shawn Mankad guided me through many interesting data science projects.

I would also like to acknowledge M. Cristina Garcia and Mr. William Schildknecht for their help and support on administrative issues.

I would like to acknowledge the financial support received from the Loccioni, and the Department of Mathematics.

Lastly, I would like to thank my family and friends. I owe my deepest thanks to my family for their love, consideration, and great confidence in me throughout all the years. Words cannot express the gratitude I owe them. I would like to thank my girlfriend Yue Xu. With her love and passion, I complete this milestone. My friends and fellow classmates at the Mathematics Department have enriched my graduate life in many ways and deserve a special mention. I would like to thank Ying Han, Hechao Sun, Cheng Jie, Ye Chen and Jinhang Xue for all their help during my course of study at Maryland.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

NPMLE	Nonparametric Maximum Likelihood Estimate
DRM	Density Ratio Model
EP	Empirical Method
EVD	Extreme Value Distribution
EVT	Extreme Value Theory
GEV	Generalized Extreme Value Distribution
GPD	Generalized Pareto Distribution
IG	Inverse Gaussian
MAE	Mean Absolute Error
OSF	Out of Sample Fusion
POT	Peaks Over Threshold
BM	Block Maxima
ROSF	Repeated Out of Sample Fusion
IM	Iterative Method
SP	Semiparametric Method

Chapter 1: The Density Ratio Model

1.1 Biased Sampling Models

The origin of the Density Ratio Model (DRM) can be traced back at least to Vardi's length - biased sampling models [36]. In Vardi's study, the length of an object is assumed to be distributed according to the cdf G, and the selection probability for any particular object is proportional to its length. Then the distribution of the length of sampled objects is given by the following model,

$$F(y) = \frac{1}{\mu} \int_0^y x dG(x), \quad y \geq 0$$

where $\mu = \int_0^\infty x dG(x) < \infty$ is the normalization constant. Here the cdf G is unknown. The cdf F is the length-biased distribution corresponding to G. It can be seen as a weighted version of G in terms of the weight function x. Gilbert et al. (1999) [18] later generalized the two sample model to allow for $s+1$ different biased samples:

$$F_i(y) = W_i(G)^{-1} \int_{-\infty}^y w_i(x) dG(x), \quad i = 1, \ldots, s$$

where the w_i's are given nonnegative selection bias weight function and

$$W_i(G) = \int_{-\infty}^\infty w_i(x) dG(x)$$

A simple way to estimate G is to use the empirical distribution of the reference sam-

ple X_0 only. This approach ignores the rest s samples. Vardi (1985) [35] developed a methodology for obtaining a nonparametric maximum likelihood estimate (NPMLE) by using all the $n = n_0 + n_1 + \cdots + n_s$ observations from the $s+1$ samples. In Vardi (1985) [35], the weight functions were assumed completely known. However, in the real data application, this assumption is unrealistic. To address this problem, we can assume that the weight function comes from a parametric family. In this situation, we need to estimate two parts in the model. First is the unknown reference distribution G and second is the parameters in the weight function. These types of models are called biased sampling semiparametric models, one typical example of these model is the logistic regression model in case-control studies.

1.2 Logistic Regression Model in Case-Control Studies

Case-control studies are common used methods to study risk factors in epidemiological observational study. Logistic regression models is the most commonly used models. Let $D = 0$ be the control, $D = 1$ be the case, $\mathbf{x} = (x_1, \ldots, x_p)$ be the regression covariates, and let $P(D = i \mid \boldsymbol{x})$ denote the probability that individual with covariates $\boldsymbol{x}$ develops disease $D = i$. The logistic regression model can be expressed as:

$$P(D = i \mid \boldsymbol{x}) = \frac{\exp(\alpha_i + \boldsymbol{\beta}_i' \boldsymbol{x})}{1 + \exp(\alpha_1 + \boldsymbol{\beta}_1' \boldsymbol{x})}, \qquad i = 0, 1 \tag{1.1}$$

let $p(\boldsymbol{x})$ be the marginal distribution of $\boldsymbol{x}$, and let $\pi_i = P(D = i)$ (note that $\sum_{i=0}^{1} \pi_i = 1$). Then by Bayes' Rule, we have:

$$P(\boldsymbol{x} \mid D = i) = \frac{P(D = i \mid \boldsymbol{x}) p(\boldsymbol{x})}{\pi_i}, \quad i = 0, 1$$

Therefore,

$$\frac{P(\boldsymbol{x} \mid D=1)}{P(\boldsymbol{x} \mid D=0)} = \frac{\pi_0}{\pi_1} \frac{P(D=1 \mid \boldsymbol{x})}{P(D=0 \mid \boldsymbol{x})} \tag{1.2}$$

By substituting (1.1) into (1.2) and letting $\alpha_0 = \beta_0 = 0$, we get the density ratio model:

$$\frac{P(\boldsymbol{x} \mid D=1)}{P(\boldsymbol{x} \mid D=0)} = \exp(\alpha_1^* + \boldsymbol{\beta}_1' \boldsymbol{x})$$

where $\alpha_1^* = \log(\pi_0/\pi_1) + \alpha_1$.

If we let $g_i(\boldsymbol{x})$ denote the conditional density function $P(\boldsymbol{x} \mid D = i)$, $i = 0, 1$, then we can rewrite the previous formula as:

$$g_1(\boldsymbol{x}) = \exp(\alpha_1^* + \boldsymbol{\beta}_1' \boldsymbol{x}) g_0(\boldsymbol{x})$$

hence the case distribution becomes a weighted version of the control distribution. This is a tilt density ratio model. The exponential function is the weight, $\boldsymbol{x}$ is called the distortion function, and the function $g_0(\boldsymbol{x})$ is regarded as the density of the reference (control) sample.

1.3 Density Ratio Models

Motivated by either biased sampling models or case-control studies, density ratio models were developed and elaborated in Qin and Lawless (1994) [32], Qin and Zhang (1997) [31], Fokianos et al. (2001) [11], Kedem et al. (2008) [20], Voulgaraki et al. (2012) [34], Zhou (2013) [39] , Pan (2016) [29], Yu (2017) [37]. For the two-sample case:

$$
\begin{aligned}
X_0 &= (x_{01}, \ldots, x_{0n_0})' \sim g_0(x) \\
X_1 &= (x_{11}, \ldots, x_{1n_1})' \sim g_1(x),
\end{aligned}
$$

so that the density ratio model is:

$$
\frac{g_1(x)}{g_0(x)} = e^{\alpha + \boldsymbol{\beta}' \boldsymbol{h}(x)} \tag{1.3}
$$

where we call $\boldsymbol{h}(x)$ the tilt function, which can be regarded as a distortion of the X_1's pdf relative to the reference X_0's pdf. Now let's consider two exponential cases.

1.3.1 When both $g_0(x, \boldsymbol{\theta})$ and $g_1(x, \boldsymbol{\theta})$ come from the same exponential family $\{P_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}^k$:

$$
P_\theta = p(x, \boldsymbol{\theta}) = d(\boldsymbol{\theta}) S(x) \exp\left[\sum_{j=1}^{k} C_j(\boldsymbol{\theta}) T_j(x)\right], \quad x \in \mathcal{X} \subset \mathbb{R}^q
$$

where $C_1, \ldots, C_k$ and d are real-valued functions of $\boldsymbol{\theta}$, and the real valued functions $T_1, \ldots, T_k$ and S have their supports in $\mathbb{R}^q$. Then:

$$
\begin{aligned}
\frac{g_1(x)}{g_0(x)} &= \frac{d(\boldsymbol{\theta}_1)}{d(\boldsymbol{\theta}_0)} \exp\left\{\sum_{j=1}^{k} \big[C_j(\boldsymbol{\theta}_1) - C_j(\boldsymbol{\theta}_0)\big] T_j(x)\right\} \\
&= \exp\left\{\sum_{j=1}^{k} \big[C_j(\boldsymbol{\theta}_1) - C_j(\boldsymbol{\theta}_0)\big] T_j(x) + \log \frac{d(\boldsymbol{\theta}_1)}{d(\boldsymbol{\theta}_0)}\right\} \\
&= \exp\Big\{\alpha + \boldsymbol{\beta}' \mathbf{h}(x)\Big\}
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha &= \log \frac{d(\boldsymbol{\theta}_1)}{d(\boldsymbol{\theta}_0)} \\
\boldsymbol{\beta} &= (C_1(\boldsymbol{\theta}_1) - C_1(\boldsymbol{\theta}_0), \ldots, C_j(\boldsymbol{\theta}_1) - C_j(\boldsymbol{\theta}_0), \ldots, C_k(\boldsymbol{\theta}_1) - C_k(\boldsymbol{\theta}_0))', \quad j = 1, \ldots k \\
\mathbf{h}(x) &= (T_1(x), \ldots, T_j(x), \ldots, T_k(x))', \quad j = 1, \ldots, k
\end{aligned}
$$

Here is a list of one to one correspondences between the tilt functions $h(t)$ and common pdf's:

$h(t)$	distribution
$h(t) = t$	$g(x) \sim \exp(\lambda)$
$\mathbf{h}(t) = (t, t^2)'$	$g(x) \sim \mathrm{N}(\mu, \sigma^2)$
$\mathbf{h}(t) = (t, \log(t))'$	$g(x) \sim \Gamma(k, \lambda)$
$\mathbf{h}(t) = (\log(t), \log(1-t))'$	$g(x) \sim \mathrm{Beta}(\alpha, \beta)$

1.3.2 When $g_0(x, \boldsymbol{\theta})$ and $g_1(x, \boldsymbol{\theta})$ come from different exponential families with the same support:

$$\begin{aligned}
\frac{g_1(x)}{g_0(x)} &= \frac{d_1(\boldsymbol{\theta}_1)S_1(x)}{d_0(\boldsymbol{\theta}_0)S_0(x)} \exp\left\{\sum_{j=1}^{k}\left[C_{1j}(\boldsymbol{\theta}_1)T_{1j}(x) - C_{0j}(\boldsymbol{\theta}_0)T_{0j}(x)\right]\right\} \\
&= \exp\left\{\sum_{j=1}^{k}\left[C_{1j}(\boldsymbol{\theta}_1)T_{1j}(x) - C_{0j}(\boldsymbol{\theta}_0)T_{0j}(x)\right] + \log\frac{d_1(\boldsymbol{\theta}_1)}{d_0(\boldsymbol{\theta}_0)} + \log\frac{S_1(x)}{S_0(x)}\right\} \\
&= \exp\{\alpha + \phi(x, \boldsymbol{\beta})\}
\end{aligned}$$

where

$$\begin{aligned}
\alpha &= \log\frac{d(\boldsymbol{\theta}_1)}{d(\boldsymbol{\theta}_0)} \\
\phi(x, \boldsymbol{\beta}) &= \sum_{j=1}^{k}\left[C_{1j}(\boldsymbol{\theta}_1) \cdot T_{1j}(x) - C_{0j}(\boldsymbol{\theta}_0) \cdot T_{0j}(x)\right] + \log\frac{S_1(x)}{S_0(x)}
\end{aligned}$$

1.4 Semiparametric Density Ratio Models

The semiparametric density ratio model establishes relationships between a reference distribution and its tilted versions. The multiple sample semiparametric

density ratio model describes the following $m+1$ independent samples:

$$\begin{aligned} X_0 &= (x_{01}, \ldots, x_{0n_0})' \sim g(x) \\ X_1 &= (x_{11}, \ldots, x_{1n_1})' \sim g_1(x) \\ &\vdots \\ X_m &= (x_{m1}, \ldots, x_{mn_m})' \sim g_m(x) \end{aligned}$$

where $g_j(x)$ is the probability density of the jth sample of size n_j. We call X_0 the reference sample. Its cumulative distribution $G(x)$ is assumed to be unknown. To estimate g and G, we assume there are additional samples from related distributions.

The density ratio model assumes that the reference distribution $g(x)$ and its tilted versions $g_j(x)$ are related by the ratios,

$$\begin{aligned} \frac{g_1(x)}{g(x)} &= \exp(\alpha_1 + \boldsymbol{\beta}_1' \boldsymbol{h}(x)) \\ &\vdots \\ \frac{g_m(x)}{g(x)} &= \exp(\alpha_m + \boldsymbol{\beta}_m' \boldsymbol{h}(x)) \end{aligned} \tag{1.4}$$

This gives the tilt model:

$$g_j(x) = e^{\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x)} g(x), \quad j = 1, \ldots, m$$

where the $\boldsymbol{\beta}_j$ are $p \times 1$ parameter vectors, the α_j are scalar parameters, and $\boldsymbol{h}(x)$ is a vector valued distortion or tilt function. The probability densities g, g_1, ... , g_m and the parameters α's and $\boldsymbol{\beta}$'s are unknown, $\boldsymbol{h}$ is assumed to be a known function. The relationship (1.4), called the density ratio model. The density ratio model allows semiparametric inference about all the parameters and distributions from the fused $m+1$ sample,

$$\mathbf{t} = (\mathrm{t}_1, ..., \mathrm{t}_\mathrm{n})' = (\mathrm{X}_0', \mathrm{X}_1', ..., \mathrm{X}_\mathrm{m}')' \tag{1.5}$$

of size $n = n_0 + n_1 + ... + n_m$. Since $n_0 < n$, the reference G, under (1.4), is estimated with all the data. For a thoroughly explanation of the semiparametric inference under (1.4), see, for example, Fokianos et al. (2001) [11], Fokianos (2004) [12], Lu (2007) [26], and Qin and Zhang (1997) [31]. A general reference is the recent book by Kedem et al. (2017) [21].

1.5 Estimation

Maximum likelihood estimates for all the parameters and $G(x)$ can be obtained by maximizing the empirical likelihood over the class of step cumulative distribution functions with jumps at the observed values $t_1, ..., t_n$. See Owen (2001) [28]. The estimate of the reference distribution function G is supported at all the n observed values $t_1, ..., t_n$ and not just at the n_0 values from the reference sample X_0. Thus, if we let $p_i = dG(t_i)$ be the mass at t_i, for $i = 1, ..., n$, the empirical likelihood becomes

$$\mathcal{L}(\boldsymbol{\theta},\mathrm{G}) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp(\alpha_1 + \boldsymbol{\beta}_1' \boldsymbol{h}(x_{1j})) \times \cdots \times \prod_{j=1}^{n_m} \exp(\alpha_m + \boldsymbol{\beta}_m' \boldsymbol{h}(x_{mj})), \tag{1.6}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_m')'$, and $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$. We maximize $\mathcal{L}(\boldsymbol{\theta},\mathrm{G})$ subject to the constraints $\sum_{i=1}^{n} p_i = 1$ and

$$\sum_{i=1}^{n} p_i[w_1(t_i) - 1] = 0, ..., \sum_{i=1}^{n} p_i[w_m(t_i) - 1] = 0$$

where $w_j(x) = \exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x))$, $j = 1, ..., m$. We obtain the desired estimates through the method of Lagrange multipliers. First we set up the objective function

$$\log \mathcal{L}(\theta, G) - \lambda_0(1 - \sum_{i=1}^{n} p_i) - \lambda_1 \sum_{i=1}^{n} p_i[w_1(t_i) - 1] - \cdots - \lambda_m \sum_{i=1}^{n} p_i[w_m(t_i) - 1],$$

to obtain $\lambda_0 = n$ and $\lambda_j = n_j \quad j = 1, \ldots, m$ and

$$p_i = \frac{1}{n_0} \cdot \frac{1}{1 + \rho_1 w_1(t_i) + \cdots + \rho_m w_m(t_i)},$$

where $\rho_j = n_j/n_0, j = 1, ..., m$. Next we substitute the p_i's into $\mathcal{L}(\boldsymbol{\theta}$,G) to get the profile log likelihood as a function of $\boldsymbol{\theta}$ only:

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= -n \log n_0 - \sum_{i=1}^{n} \log[1 + \rho_1 w_1(t_i) + \cdots + \rho_m w_m(t_i)] \\
&+ \sum_{j=1}^{n_1} (\alpha_1 + \boldsymbol{\beta}_1' \boldsymbol{h}(x_{1j})) + \cdots \\
&+ \sum_{j=1}^{n_m} (\alpha_m + \boldsymbol{\beta}_m' \boldsymbol{h}(x_{mj})).
\end{aligned}$$

Then, we differentiate the objective function $\log \ell$ with respect to the α_i and β_i to get the score equations:

$$\begin{aligned}
\frac{\partial l}{\partial \alpha_j} &= -\sum_{i=1}^{n} \frac{\rho_j w_j(t_i)}{1 + \rho_1 w_1(t_i) + \cdots + \rho_m w_m(t_i)} + n_j = 0 \\
\frac{\partial l}{\partial \boldsymbol{\beta}_j} &= -\sum_{i=1}^{n} \frac{\rho_j \boldsymbol{h}(t_i) w_j(t_i)}{1 + \rho_1 w_1(t_i) + \cdots + \rho_m w_m(t_i)} + \sum_{i=1}^{n_j} \boldsymbol{h}(x_{ji}) = \mathbf{0}
\end{aligned}$$

The solution of the score equations, which is found numerically, gives the maximum likelihood estimators $\hat{\alpha}, \hat{\boldsymbol{\beta}}$ and consequently by substitution:

$$\hat{p}_i = \frac{1}{n_0} \cdot \frac{1}{1 + \sum_{j=1}^{m} \rho_j \exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}'_j \mathbf{h}(t_i))}. \tag{1.7}$$

In particular, the maximum likelihood estimate $\hat{G}$ of G is given in (1.8) for relative sample sizes $\rho_j = n_j/n_0$:

$$\hat{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n} \frac{I(t_i \le t)}{1 + \rho_1 \hat{w}_1(t_i) + ... + \rho_m \hat{w}_m(t_i)} \tag{1.8}$$

where $\hat{w}_j(x) = \exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}'_j \boldsymbol{h}(x))$, $j = 1, ..., m$, and $I(t_i \le t)$ equals one for $t_i \le t$ and is zero otherwise. Similarly, $\hat{G}_j$ can be estimated by accumulating $\exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}'_j \boldsymbol{h}(t_i)) dG(t_i)$.

1.6 Asymptotic Results for $\hat{\boldsymbol{\theta}}$ and $\hat{G}$

The asymptotic behavior of the parameter estimators $\hat{\alpha}, \hat{\boldsymbol{\beta}}$ and the estimator for the reference cdf $\hat{G}$ are studied in Qin and Zhang (1997) [31] and Zhang (2000) [38] for the two sample case. The multiple sample case was discussed by Lu (2007) [26] using the same strategy.

1.6.1 Asymptotic Theory for $\hat{\boldsymbol{\theta}}$

Let $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0)$ be the true value of $(\alpha, \boldsymbol{\beta})$. Then under the density ratio model, as $n \to \infty$,

$$\sqrt{n} \begin{pmatrix} \hat{\alpha} - \alpha_0 \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \Rightarrow \mathrm{N}(\mathbf{0}, \mathbf{S^{-1}V S^{-1}})$$

where:

$$\mathbf{V} \equiv \mathrm{Var}\left[\frac{1}{\sqrt{n}} \nabla \ell(\alpha, \boldsymbol{\beta})\right], \quad \mathbf{S} \equiv \lim_{n\to\infty}\left[-\frac{1}{n} \nabla\nabla' \ell(\alpha, \boldsymbol{\beta})\right]$$

Note that $\mathbf{V}, \mathbf{S}$ are $(1+p)m \times (1+p)m$ matrices.

The strong consistency of $\hat{\boldsymbol{\theta}}$ as the estimator of the true parameter $\boldsymbol{\theta}_0$ has been established in Lu (2007) [26], where more details are given.

1.6.2 Asymptotic Theory for $\hat{G}$

The multiple sample asymptotic behavior of $\hat{G}$ was also obtained by Lu (2007) [26], from which we obtain semiparametric (SP) confidence intervals by using the covariance matrix given below.

$$
\begin{aligned}
A_j(t) &= \int \frac{w_j(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y) \\
B_j(t) &= \int \frac{w_j(y) h(y) I(y \le t)}{\sum_{k=0}^{m} \rho_k w_k(y)} dG(y)
\end{aligned}
$$

$$
\bar{A}(t) = (A_1(t), ..., A_m(t))' \quad , \quad \bar{B}(t) = (B_1'(t), ..., B_m'(t))'
$$

$$
\boldsymbol{\rho} = \mathrm{diag}\{\rho_1, \ldots, \rho_m\}_{m \times m}, \quad \mathbf{1}_p = (1, ..., 1)'
$$

Then the asymptotic distribution of $\hat{G}(t)$ for $m \ge 1$ is given by the following two theorems, assuming that all moments with respect to the reference distribution are finite.

Theorem 1.1. *The process $\sqrt{n}(\hat{G}(t) - \tilde{G}(t))$ converges weakly to a zero-mean Gaussian process W with continuous sample paths in the space of real right continuous functions, and the covariance matrix is determined by*

$$
\begin{aligned}
Cov\left\{\sqrt{n}(\hat{G}(t) - \tilde{G}(t)),\ \sqrt{n}(\hat{G}(s) - \tilde{G}(s))\right\} &= \sum_{k=0}^{m} \rho_k \sum_{j=1}^{m} \rho_j A_j(t \wedge s)) \\
&\quad - \left(\bar{A}'(t)\boldsymbol{\rho}, \bar{B}'(t)(\boldsymbol{\rho} \otimes \mathbf{1}_p)\right) S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(s) \\ (\boldsymbol{\rho} \otimes \mathbf{1}_p)\bar{B}(s) \end{pmatrix}
\end{aligned}
$$

Theorem 1.2. *The process $\sqrt{n}(\hat{G}(t) - G(t))$ converges weakly to a zero-mean Gaussian process in the space of real right continuous functions, with covariance matrix given by*

$$
\begin{aligned}
Cov\left\{\sqrt{n}(\hat{G}(t) - G(t)),\ \sqrt{n}(\hat{G}(s) - G(s))\right\} &= (\sum_{k=0}^{m} \rho_k)(G(t \wedge s) - G(t)G(s) - \sum_{j=1}^{m} \rho_j A_j(t \wedge s)) \\
&\quad + \left(\bar{A}'(s)\boldsymbol{\rho}, \bar{B}'(s)(\boldsymbol{\rho} \otimes \mathbf{1}_p)\right) S^{-1} \begin{pmatrix} \boldsymbol{\rho}\bar{A}(t) \\ (\boldsymbol{\rho} \otimes \mathbf{1}_p)\bar{B}(t) \end{pmatrix}
\end{aligned}
$$

where $\mathbf{1}_p$ is the $p \times p$ identity matrix, $\tilde{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} I[x_{0i} < t]$ is the empirical distribution of the reference sample X_0 only, and $\otimes$ denotes the Kronecker product.

The complete derivation of the theorems can be found in Lu (2007) [26]. The immediate application of Theorem 1.1 is in the construction of pointwise symmetric confidence intervals for $G(t)$ for any given t.

1.7 Goodness of Fit Tests

Goodness of fit tests are needed to justify the applicability of the density ratio model. Let $\hat{G}(t)$ be the estimated reference cdf and $\tilde{G}(t)$ be the empirical cdf of the reference sample. Most goodness of fit tests measure the discrepancy between $\hat{G}(t)$ and $\tilde{G}(t)$. A simple graphical method is to plot $\hat{G}(t)$ versus $\tilde{G}(t)$. See Voulgaraki et al. (2012) [34]. A numerical method is proposed in Qin and Zhang (1997) [31]. Define the difference between $\hat{G}(t)$ and $\tilde{G}(t)$ as:

$$\Delta_n(t) = \sqrt{n}\,|\hat{G} - \tilde{G}|, \qquad \Delta_n = \sup_{-\infty<t<\infty} \Delta_n(t)$$

then Δ_n can be used to measure the departure from the assumption of the semi-parametric density ratio model. Theorem 1.1 shows that $\sqrt{n}(\hat{G}(t)-\tilde{G}(t))$ converges weakly to a Gaussian process W. Let w_α denote the α-quantile of the distribution of $\sup_{-\infty<t<\infty}|W(t)|$. By Theorem 1.1,

$$\begin{aligned}\lim_{n\to\infty} P(\Delta_n \geq w_{1-\alpha}) &= \lim_{n\to\infty} P(\sup_{-\infty<t<\infty} \sqrt{n}\,|\hat{G} - \tilde{G}| \geq w_{1-\alpha}) \\ &= P(\sup_{-\infty<t<\infty} \sqrt{n}\,|W(t)| \geq w_{1-\alpha}) = \alpha\end{aligned}$$

The density ratio model is rejected at level α if $\Delta_n \geq w_{1-\alpha}$. However, there is no analytic expression available for the distribution of the supremum of a Gaussian process $W(t)$ and its corresponding quantile function. Ofter, a bootstrap procedure is applied to simulate the distribution of $\sup_{-\infty<t<\infty}|W(t)|$ and its quantiles.

Yu (2017) [37] discussed a goodness-of-fit test based on the discrepancy between $\hat{g}$ estimated under the density ratio model from the entire fused data $\boldsymbol{t}$, and $\tilde{g}$ estimated from the reference sample $\boldsymbol{x}_0$ only. For a given kernel K with a fixed bandwidth b, we can construct density estimators as follows:

$$\begin{aligned} \hat{g}(t) &= \int K(t-y)d\hat{G}(y) \\ \tilde{g}(t) &= \int K(t-y)d\tilde{G}(y). \end{aligned}$$

Yu (2017) [37] defined a new test statistic in terms of the Hellinger distance

$$I_n = nb\int_{-L}^{L}(\sqrt{\hat{g}(t)}-\sqrt{\tilde{g}(t)})^2dt$$

where $[-L, L]$ is a closed and bounded interval. A detailed theoretical derivation of the asymptotic distribution of I_n can be found in Yu (2017) [37].

Chapter 2: Semiparametric Quality Control

This chapter discusses an application of the Density Ratio Model to mechanical quality control. Two real data problems will be included in this chapter.

2.1 Introduction

Acceleration data obtained from machine vibration are used routinely in quality control, particularly in deciding "normal" versus "faulty" or "good" versus "bad" mechanical systems such as electric motors and car engines, or mechanical components such as ball bearings and tires (Concettoni et al. 2012, Cristali et al. 2006, Goyal and Pabla 2016) [4, 5, 17]. The purpose of this chapter is to illustrate the semiparametric statistical method applied in the analysis of accelerometer data for the purpose of quality control. The method is based on *fusion* of records from several sampled signatures, be they normal or faulty (Kedem et al. 2017) [21], and seems to be highly effective.

Given two or more vibration signatures, the idea is to obtain a great data reduction and use only representative random samples from each signature. Since in general the original signals are much larger than any sample, the analysis can be repeated by redrawing many additional random samples for quality assurance

purposes.

An advantage of the method is the fact that it is based on probability distributions and not spectral quantities. Spectral methods are in general very effective; however in extreme cases it is possible for signals to have the exact same spectrum but very different distributions. This is exemplified by two first order autoregressive processes

$$z_t = \phi z_{t-1} + \epsilon_t,$$

one with normal noise, $\epsilon_t \sim N(0,1)$, and the other with normalized t-distributed noise, $\epsilon_t \sim t_{(3)}/\sqrt{3}$, and both with the same parameter, say $\phi = 0.8$. In that case the autocorrelation functions, and hence also the corresponding spectral densities, are identical but the marginal distributions are markedly different. This is illustrated in Figure 2.1 in terms of estimated quantities.

For illustrative purposes, we shall deal here with ball bearing and electric motor accelerometer data. The semiparametric method works as follows. A benchmark signature, usually "good", is chosen and is sampled randomly to produce a *reference sample*. This reference random sample is distributed according to an *unknown reference distribution*. In the present application, once a reference random sample is obtained from some unknown reference probability distribution, the method is very sensitive to deviations from the reference distribution as expressed by very low p-values, essentially close to zero, meaning a *different statistical behavior*. On the other hand, when equality of distributions (equidistribution) occurs, the p-values are unusually high, meaning *similar statistical behavior*. Hence, the method is an

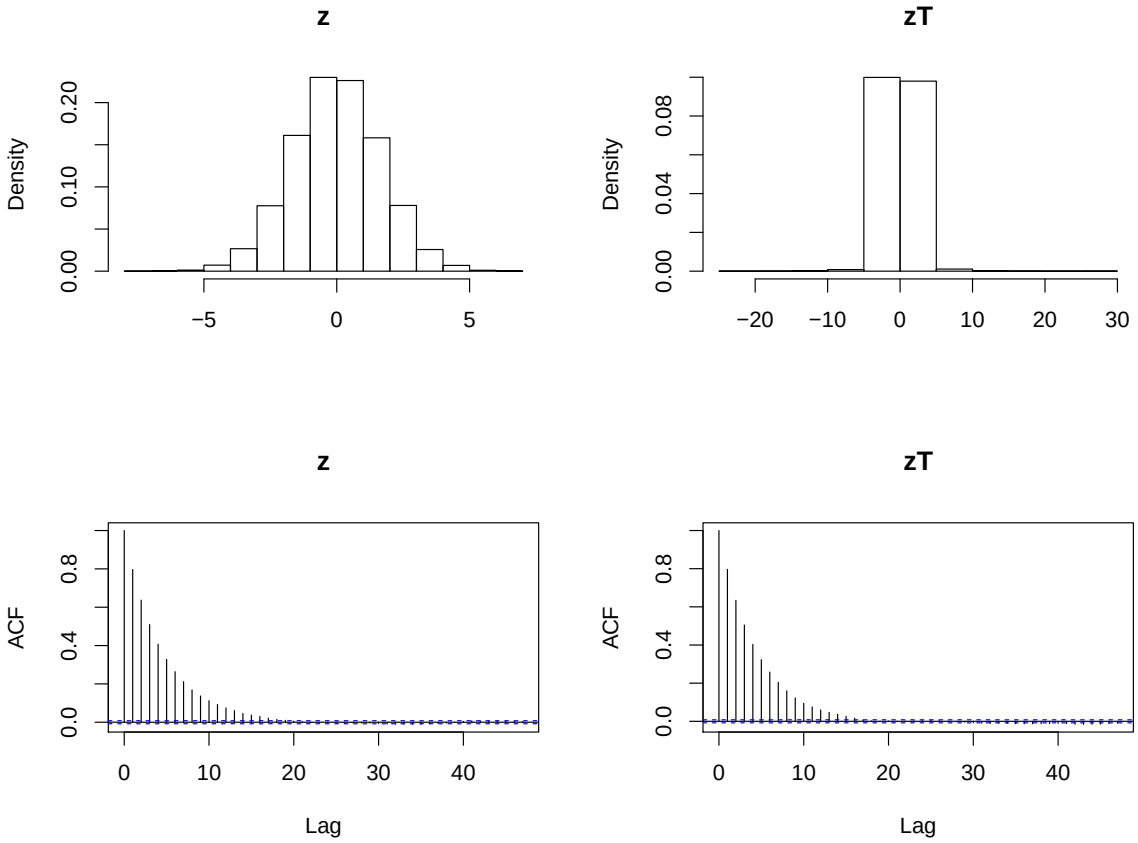

Figure 2.1: Different histograms corresponding to identical autocorrelations from a first order autoregressive process with normal noise (z) and t-noise (zT).

additional tool useful in detecting abnormalities in quality control.

Detecting differences among probability distributions can be approached via the so called *density ratio model* along with the appropriate semiparametric statistical inference discussed briefly in the next section and in Chapter 1.

2.2 A Semiparametric Method

The density ratio model allows statistical inference about unknown probability distributions representing many sources by fusing samples obtained from each source. The only assumption is a *connection* between the distributions.

We follow the construction proposed in a recent general reference (Kedem, et al. (2017) [21]). An earlier reference along the same lines in terms of data from two radars is in Kedem et al. (2004) [22]. Additional related references are Fokianos et al. (2001) [11], Gilbert at al. (1999) [18], Qin and Zhang (1997) [31], and Vardi (1982,1985) [35,36]. We summarize the essence of the method in what follows. More details are given in Chapter 1.

Assume there are $m+1$ data sources from which we obtain, respectively, random samples $\mathbf{x_0}, \boldsymbol{x_1}, ..., \boldsymbol{x_m}$, where $\boldsymbol{x}_j$ is of size n_j and is governed by a probability density $g_j(x)$. This is expressed as,

$$x_{ji} \sim g_j(x), \quad j = 0, 1, ..., m, \quad i = 1, ..., n_j,$$

and we let $g_0(x) = g(x)$ be the *reference* probability density function (pdf). The samples are fused or combined in a long vector of size $n = n_0 + n_1 + \cdots + n_m$,

$$\boldsymbol{t} = (t_1, ..., t_n)' \equiv (\boldsymbol{x}_0', \boldsymbol{x}_1', ..., \boldsymbol{x}_m').$$

Our semiparametric statistical inference uses the entire fused data $\boldsymbol{t}$ in the estimation of the probability densities ("distortions") $g_1, ..., g_m$ and the reference $g_0 = g$. Thus, for example, g_1 is estimated from the entire fused data $\boldsymbol{t}$ and not just from $\boldsymbol{x}_1$, and g_2 is estimated from the entire fused data $\boldsymbol{t}$ and not just from $\boldsymbol{x}_2$, and so on. Since the fused data are larger than any individual sample, this gives more precise statistical inference than any inference based on any particular sample.

For a given tilt function $\boldsymbol{h}(x)$ (which could be a vector or a scalar), it is assumed that the m distortions of the reference g satisfy the *density ratio model*,

$$g_j(x) = \exp\{\alpha_j + \boldsymbol{\beta}_j^{'}\boldsymbol{h}(x)\}g(x), \quad j = 1, ..., m. \tag{2.1}$$

We wish to test hypotheses about the p-dimensional parameters $\boldsymbol{\beta}_j$, and in particular test distribution equality (equidistribution or equal statistical behavior),

$$H_0 : \boldsymbol{\beta}_1 = ... = \boldsymbol{\beta}_m = \mathbf{0}.$$

Notice that $\boldsymbol{\beta_j} = \mathbf{0}$ implies $\alpha_j = 0$, in which case $g_j = g$. Hence, under H_0 all the densities "agree", $g = g_0 = g_1 = \cdots = g_m$, and we have equidistribution. That is, all sources behave alike statistically. To test the hypothesis H_0 of equidistribution we use the likelihood ratio test discussed briefly in what follows.

Following the development in Qin and Zhang (1997) [31], let G be the reference cumulative distribution function corresponding to g and let $p_i = dG(t_i)$, $i = 1, ..., n$. Then the semiparametric likelihood becomes,

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, p_1, ..., p_n) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} w_1(x_{1j}) \cdots \prod_{j=1}^{n_m} w_m(x_{mj}) \tag{2.2}$$

where $w_j(t) = \exp\{\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(t)\}$. The maximum likelihood estimator can be obtained from equation (1.7) and the estimate of the reference cdf G from the entire fused data $\boldsymbol{t}$ is

$$\hat{G}(x) = \sum_{i=1}^{n} \hat{p}_i I(t_i \leq x) \tag{2.3}$$

where $I(B)$ denotes the indicator function of event B. The estimate of the reference distribution G from the reference sample only, $\boldsymbol{x}_0 = (x_{01}, \ldots, x_{0n_0})$, is the empirical distribution $\tilde{G}$,

$$\tilde{G}(t) = \frac{1}{n_0} \sum_{i=1}^{n_0} I(x_{0i} \leq t). \tag{2.4}$$

If ℓ denotes the resulting log-likelihood, which is now a function of α's and $\boldsymbol{\beta}$'s, then the likelihood ratio test for testing H_0 uses the statistic

$$LR \equiv -2[\ell(\mathbf{0}, \mathbf{0}) - \ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})] \tag{2.5}$$

and H_0 is rejected for large values of LR, using the fact that LR is asymptotically distributed as χ^2 with pm degrees of freedom. For more details, see Kedem et al. (2017) [21].

It should be noted that the equidistribution hypothesis H_0 goes well beyond the widespread analysis of variance where the problem is to test equality of means under the normal assumption (Fokianos et al. 2001 [11]). Here, we test *distribution equality* (not just moments) and the only distributional assumption is expressed in terms of the relationships between distributions (2.1), bypassing the normal assumption. Interestingly, (2.1) also holds in the very special case when the samples are normal for any combination of means and variances.

2.3 Application to Motor Testing

The following is a generic problem which illustrates a semiparametric approach to motor testing for the purpose of quality control.

From three acceleration signals A,B,C, we sampled three independent random samples $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2$, respectively. The vibration data and their retrieval are described in detail in Concettoni et al. (2012) [5]. For illustrative purposes, we shall assume that the three signals represent three locations A,B,C, on a motor, and that signal A is a "normal" or "good" signature serving as a benchmark. The motor is deemed healthy if the hypothesis of equidistribution (see below) is accepted.

Each sample is of size 500, and the data are fused in the long vector

$$\boldsymbol{t} = (t_1, ..., t_{1500})' \equiv (\boldsymbol{x}_0', \boldsymbol{x}_1', \boldsymbol{x}_2').$$

Our original signals are far longer corresponding to about 3 seconds worth of acceleration signatures sampled at the rate of 25.6KHz. Hence, samples of size 500 bring about a huge data reduction.

The density ratio model stipulates that $\boldsymbol{x}_j \sim g_j(x)$ for $j = 0, 1, 2$, and that $g_0(x) = g(x)$, corresponding to location A ("good"), is chosen as the reference pdf. For a given tilt function $\boldsymbol{h}(x)$, there are 2 possible distortions of the reference g, namely,

$$g_j(x) = \exp\{\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x)\} g(x), \qquad j = 1, 2. \tag{2.6}$$

Accordingly, if the hypothesis of equidistribution H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{0}$ (implying $g = g_0 = g_1 = g_2$) is accepted, then the three signatures agree and the motor is considered

healthy. Observe again that when $\boldsymbol{\beta}_j = \mathbf{0}$ then also $\alpha_j = 0$.

To continue we need the tilt function $h(x)$. The goodness-of-fit test of Qin and Zhang (1997) [31] and Yu (2017) [37] applied to numerous motor signatures points to $h(x) = x$ (a scalar) as a reasonable choice and we use it here. We have discussed goodness-of-fit in Section 1.7; the results of goodness-of-fit are applied to motor data in Section 2.5.

Again, "equidistribution" means the same statistical behavior. In general, when the hypothesis H_0 of equidistribution holds true, the cumulative distribution functions (cdf's) and the corresponding probability density functions (pdf's), representing here motor behavior, are very close; see Figure 2.2. A discrepancy is observed when the hypothesis H_0 of equidistribution is rejected; see Figures 2.3 and 2.4. Thus, hypothesis testing is done here both *analytically* as well as *graphically*. In other words, we provide the quality control user both *analytical* and *graphical* means or ways for deciding "good" versus "bad".

The likelihood ratio test (2.5) applied to five different triplets of A,B,C samples, respectively, from three healthy signatures gave the following p-values:

$$0.8645, 0.7230, 0.4979, 0.4075, 0.4849,$$

the hypothesis H_0 of equidistribution is accepted quite convincingly in each case. That is, the method was applied first to three samples from locations A,B,C giving a p-value of 0.8645. The method was applied again to different A,B,C samples for which the p-value was 0.7230, and so on five times, giving consistently high p-values as we would expect from healthy signatures. Figure 2.2 shows the results

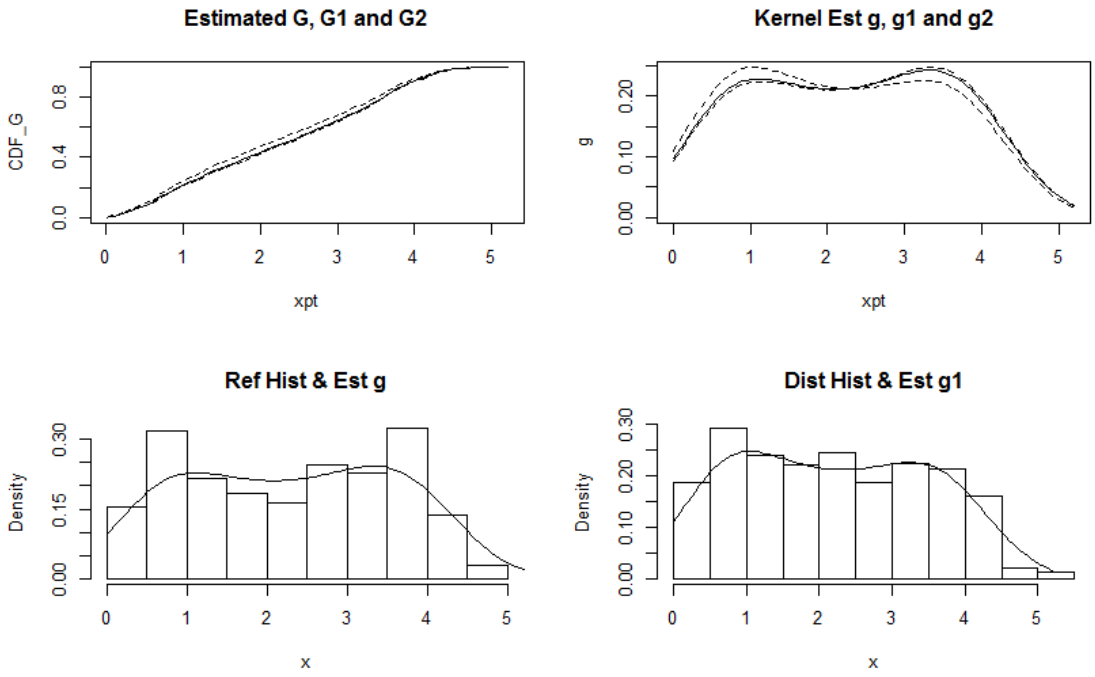

Figure 2.2: Three healthy signatures.

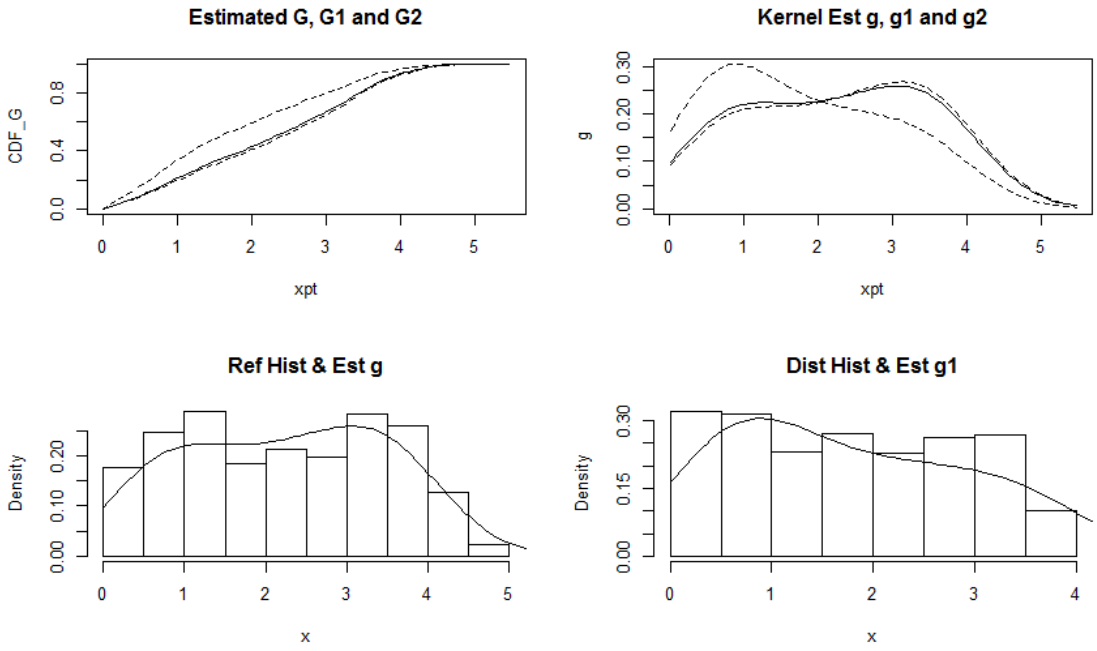

Figure 2.3: One bad signature.

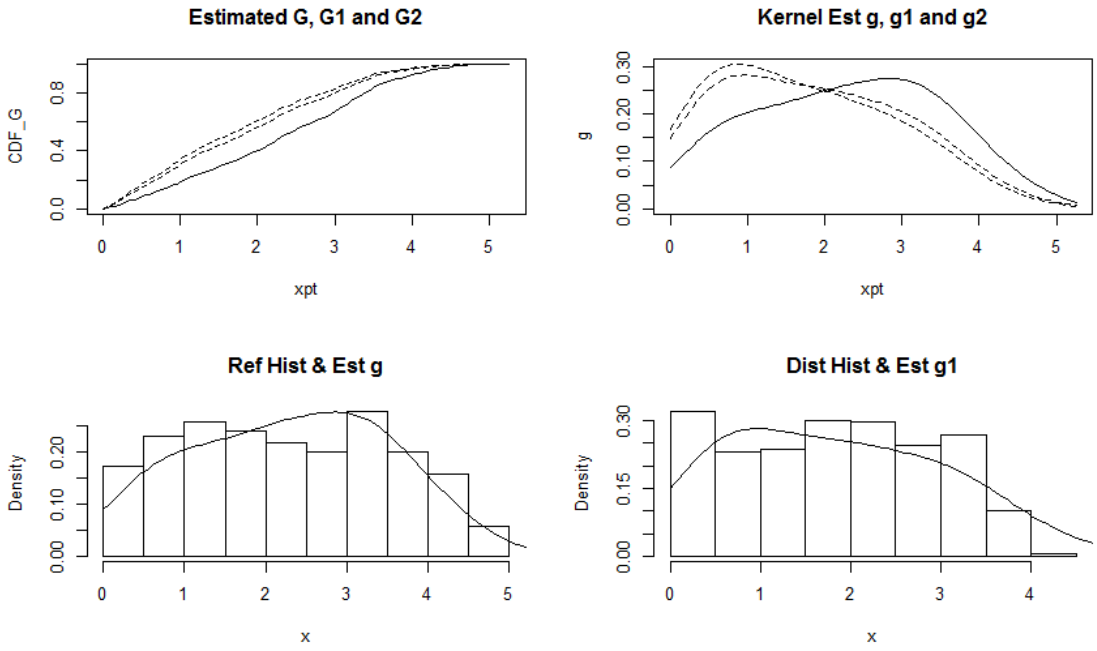

Figure 2.4: Two bad signatures.

corresponding to the p-value of 0.4849. We see that the estimated cdf's and pdf's are quite close to each other in support of H_0.

On the other hand when a single "bad" signature replaced a healthy one from location B, the p-values (again from five different trials) were reduced dramatically to

$$1.7673e^{-8}, 7.0701e^{-9}, 3.5749e^{-14}, 1.2521e^{-12}, 6.9477e^{-9},$$

and H_0 is rejected strongly in each case, pointing to the sensitivity of the method. Figure 2.3 is a graphical manifestation of the test results corresponding to a p-value of $1.2521e^{-12}$ where H_0 is rejected strongly. We observe that the first panel in Figure 2.3 is completely different from that in Figure 2.2.

When two "bad" signatures replaced healthy signatures from locations B and C, the p-values were again very small consistently,

$$8.8753e^{-10}, 1.4755e^{-13}, 2.3208e^{-7}, 5.2199e^{-11}, 2.3648e^{-14},$$

and H_0 is rejected strongly again in each trial. Figure 2.4 shows graphical results corresponding to the p-value of $2.3648e^{-14}$ where H_0 is strongly rejected. Again the first panel in Figure 2.4 is very different from the one in Figure 2.2 which depicts equidistribution.

2.4 Application to Ball Bearing Testing

The Case Western Reserve University Bearing Data Center website `https://csegroups.case.edu/bearingdatacenter/home` provides access to motor bearing test data for normal and faulty bearings.

Bearings were seeded with faults ranging from 0.007 to 0.040 inch in diameter and reinstalled into a test motor. Vibration data were recorded for motor loads of 0 to 3 horsepower (speed of 1797 to 1720 rpm). For illustration, we will compare here the following fan end (FE) bearing accelerometer data collected at 12 KHz:

GoodX098, hp=1, rpm = 1772.

BadX275, hp=1, rpm = 1772, fault diameter = 0.014 inch.

BadX279, hp=1, rpm = 1772, fault diameter = 0.007 inch.

In the present application we test "normal" versus "faulty" using two independent random samples $\boldsymbol{x}_0, \boldsymbol{x}_1$ sampled from the pairs GoodX098, BadX275 and GoodX098, BadX279, using $\boldsymbol{x}_0$ from GoodX098 as the reference sample. Again, goodness-of-fit testing discussed in Section 2.5 suggests the scalar $h(x) = x$.

In the present application each of the two samples is of size 1000, and the data are fused in the long vector

$$\boldsymbol{t} = (t_1, ..., t_{2000})' \equiv (\boldsymbol{x}_0', \boldsymbol{x}_1').$$

The original bearing signals are by far longer corresponding to at least 10 seconds worth of acceleration signatures sampled at the rate of 12KHz. Hence, we have a data reduction resulting in faster computation.

The density ratio model for $m = 1$ reduces to $\boldsymbol{x}_0 \sim g_0(x) = g(x)$, corresponding to a "good" or "normal" signature, is chosen as the reference pdf, and $\boldsymbol{x}_1 \sim g_1(x)$.

For the tilt function $h(x) = x$ we have

$$g_1(x) = \exp\{\alpha_1 + \beta_1 x\} g(x). \tag{2.7}$$

Accordingly, the hypothesis of equidistribution reduces to H_0: $\beta_1 = 0$ (implying $g_1 = g$). If the hypothesis is accepted then the second signature is "normal" as well, otherwise it is "faulty". As before, the test can be repeated multiple times with different samples to make sure the results are noncontradictory.

BadX275 versus GoodX098: The p-values from the likelihood ratio test in five different trials are very small,

$$6.8319e^{-5}, 0.00156, 2.1500e^{-5}, 5.6884e^{-9}, 2.2204e^{-15}$$

so that BadX275 is faulty with high confidence.

BadX279 versus GoodX098: Again, the p-values from the likelihood ratio test in five different trials are very small,

$$1.8902e^{-5}, 2.6968e^{-7}, 4.8627e^{-5}, 0.001204, 9.1552e^{-5}$$

so that BadX279 is also faulty with high confidence.

On the other hand, when both $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ are both from the normal vibration GoodX098, the p-values jump dramatically upward as they should,

$$0.9509, 0.9765, 0.6105, 0.9252, 0.9940.$$

As in the previous example, we see that the p-values are unusually low or unusually high, a fact which points to the potential of the semiparametric method in effective discrimination between normal and faulty bearings.

2.5 Goodness of Fit Test

The semiparametric method requires the tilt function $\boldsymbol{h}(x)$. The following goodness-of-fit tests can help in validating the usefulness of a chosen tilt function. We note that in many cases different tilt functions lead to similar hypothesis testing results; see Kedem et al. (2004) [22].

2.5.1 Δ_n statistic

Let $\hat{G}(t)$ be the estimated reference cdf from the entire fused data $\boldsymbol{t}$ under the density ratio model, and let $\tilde{G}(t)$ be the corresponding empirical cdf estimated from the reference sample $\boldsymbol{x}_0$ only where no model is assumed. Most goodness of fit tests measure the discrepancy between $\hat{G}(t)$ and $\tilde{G}(t)$, or equivalently, the discrepancy between the corresponding pdf's, in model validation. We use a useful numerical method proposed by Qin and Zhang (1997) [31]; see Section 1.7.

Define the difference between $\hat{G}(t)$ and $\tilde{G}(t)$ as:

$$\Delta_n(t) = \sqrt{n}\,|\hat{G} - \tilde{G}|, \qquad \Delta_n = \sup_{-\infty<t<\infty} \Delta_n(t).$$

Then Δ_n can be used to measure the departure from the semiparametric density ratio model with a specified $\boldsymbol{h}(x)$.

The density ratio model is rejected for large values of Δ_n, whereas small values of Δ_n lend support to the choice of $\boldsymbol{h}(x)$. Obtaining the analytical distribution of Δ_n needed for hypothesis testing is problematic. However, good approximations can be obtained by computer simulations. In the present application, this can be

readily done by sampling repeatedly from the vibration signatures, which typically are quite long.

2.5.2 I_n statistic

Another goodness-of-fit method is based on the discrepancy between $\hat{g}$ estimated under the density ratio model from the entire fused data $\boldsymbol{t}$, and $\tilde{g}$ estimated from the reference sample $\boldsymbol{x}_0$ only. A particular measure is defined in terms of the *Hellinger distance* by Yu (2017) [37]

$$I_n \;=\; nb\int_{-L}^{L}(\sqrt{\hat{g}(t)}-\sqrt{\tilde{g}(t)})^2dt$$

Again, in the present application, the distribution of I_n can be approximated by sampling repeatedly from the vibration signatures. However, this method is more involved than the previous one which is based on cdf's

2.5.3 Goodness of Fit Applied to Motor and Bearing Data

Fortunately, since the motor data is large, we can sample from it many times to simulate the distributions of Δ_n and I_n. We obtained the approximate distributions of Δ_n and I_n, shown in Figure 2.5 from 1000 applications of the density ratio model using three healthy motor signatures and $h(x) = x$. All samples were of size 500.

To demonstrate that $h(x) = x$ is a reasonable choice, we obtained from three additional healthy signatures the following results. $\Delta_{n,obs} = 0.7273$, and a rather large p-value $P(\Delta_n \geq 0.7273) = 0.673$, and $I_n = 0.0247$, and again a rather large p-value of 0.659.

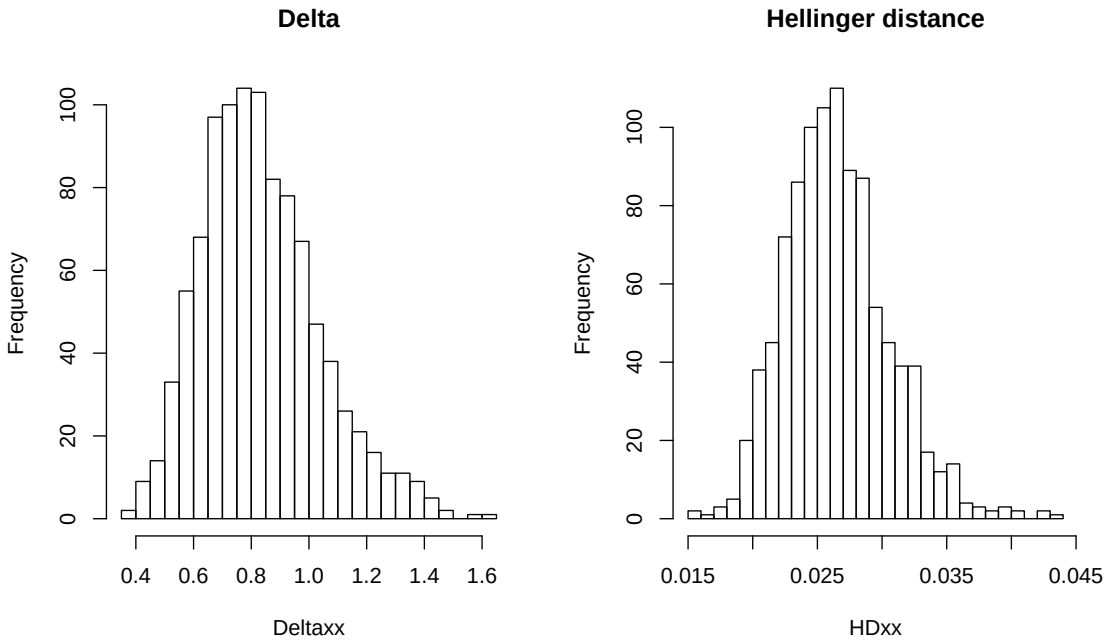

Figure 2.5: Distribution of Δ_n and I_n, motor data.

A similar analysis using the ball bearing data with sample sizes of 1000, two "normal" (good) signatures, and $h(x) = x$, gave $\Delta_{n,obs} = 0.5186$, and a rather large p-value $P(\Delta_n \geq 0.5186) = 0.835$, while $I_n = 0.000276$ giving also a rather large p-value of 0.827, again lending support to the choice of $h(x) = x$. Figure 2.6 shows the distributions of Δ_n and I_n where $h(x) = x$ for the bearing case.

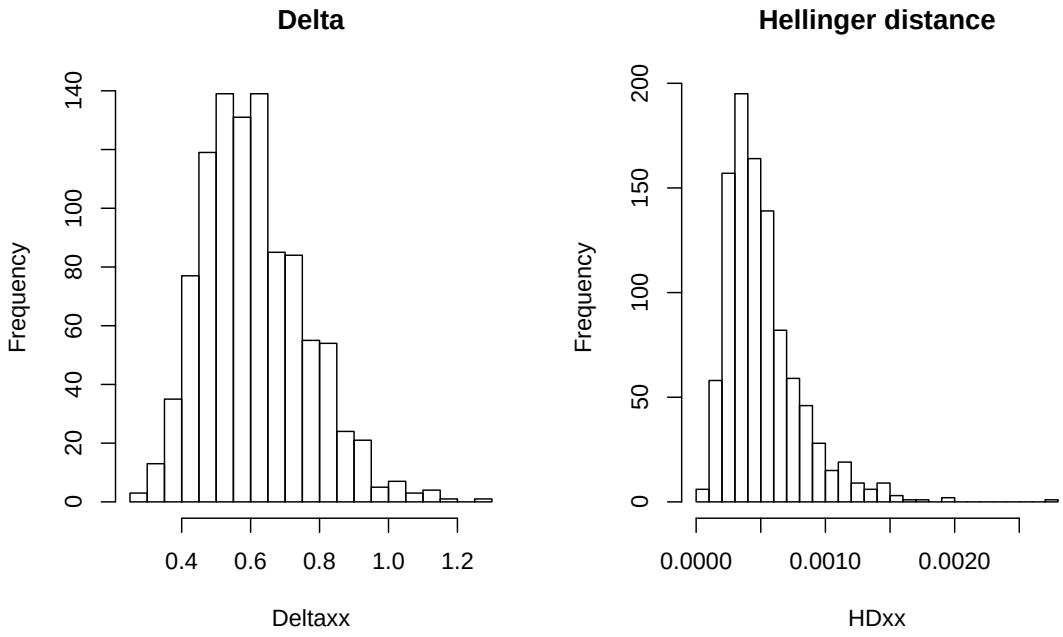

Figure 2.6: Distribution of Δ_n and I_n, ball bearing data.

2.6 The Bivariate Extension

The previous setup can be easily generalized to multivariate data. The multivariate density ratio method provides a way for determining and quantifying the differences between two or more multivariate distributions based on the joint behavior of many variables. This is very important in mechanical quality control where

the autocorrelation of mechanical time series is often very strong, and we need to include lagged variables to obtain better discriminant power. As the general multivariate case is entirely analogous to the somewhat simpler bivariate case, it is convenient to focus on the bivariate situation.

2.7 Bivariate Normal Example

Suppose we have $m+1$ two-dimensional data sets,

$$(x_{ji}, y_{ji}) \sim g_j(x,y), \quad j = 0, 1, ..., m, \quad i = 1, ..., n_j,$$

where $g_j(x,y)$ is the probability density of $N(\boldsymbol{\mu_j}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\mu_j} = \begin{pmatrix} \mu_{jx} \\ \\ \mu_{jy} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}, \quad j = 0, 1, ..., m.$$

Then, choosing $g_0(x,y)$ as a reference density we have

$$g_j(\boldsymbol{x}) = \exp[(\boldsymbol{\mu_j} - \boldsymbol{\mu_0})' \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu_j}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu_j} - \boldsymbol{\mu_0}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu_0})] g_0(\boldsymbol{x}),$$

where $\boldsymbol{x} = (x,y)'$. We see that is a special case of the general form

$$g_j(\boldsymbol{x}) = \exp\{\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(\boldsymbol{x})\} g_0(\boldsymbol{x})$$

where

$$\alpha_j = -\frac{1}{2}(\boldsymbol{\mu_j}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu_j} - \boldsymbol{\mu_0}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu_0})$$

$$\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{j1} \\ \\ \beta_{j2} \end{pmatrix} = \boldsymbol{\Sigma^{-1}}(\boldsymbol{\mu_j} - \boldsymbol{\mu_0})$$

$$h(\boldsymbol{x}) = \boldsymbol{x} = (x, y)^{'}$$

If we consider the ratio of two bivariate normal densities with unequal covariance matrices, we could use the model

$$g_j(\boldsymbol{x}) = \exp\{\alpha_j + \boldsymbol{\beta}_j^{'}\boldsymbol{h}(\boldsymbol{x})\}g_0(\boldsymbol{x})$$

where

$$\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \beta_{j3}, \beta_{j4}, \beta_{j5})^{'} \quad and \quad \boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)^{'}.$$

2.8 Likelihood Considerations

Suppose we have $m + 1$ two-dimensional data sets,

$$(x_{ji}, y_{ji}) \sim g_j(x, y), \quad j = 0, 1, ..., m, \quad i = 1, ..., n_j.$$

for a given tilt function $\boldsymbol{h}(\boldsymbol{x})$, the two dimensional density ratio model is expressed as

$$g_j(\boldsymbol{x}) = \exp\{\alpha_j + \boldsymbol{\beta}_j^{'}\boldsymbol{h}(\boldsymbol{x})\}g(\boldsymbol{x}), \quad j = 1, ..., m$$

with reference $g \equiv g_0$, and $\boldsymbol{x} = (x, y)^{'}$. The $\boldsymbol{\beta}_j$ are $p \times 1$ parameter vectors, the α_j are scalar parameters and $\boldsymbol{h}(\boldsymbol{x})$ is a vector valued distortion or tilt function.

The previous results carry over to the two-dimensional case quite readily. We begin by first defining the combined data,

$$\boldsymbol{t} = (\boldsymbol{x}_{01}^{'}, ..., \boldsymbol{x}_{0n_0}^{'}, \boldsymbol{x}_{11}^{'}, ..., \boldsymbol{x}_{1n_1}^{'}, ..., \boldsymbol{x}_{m1}^{'}, ..., \boldsymbol{x}_{mn_m}^{'})^{'} = (\boldsymbol{t}_{\boldsymbol{1}}^{'}, \boldsymbol{t}_{\boldsymbol{2}}^{'}, ..., \boldsymbol{t}_{\boldsymbol{n}}^{'})^{'}.$$

where $\boldsymbol{t_i} = (t_{ix}, t_{iy})^{'}$. Let G be the reference cumulative distribution function corresponding to g. To obtain the maximum likelihood estimator of $G(x, y)$, we optimize

over the class of two dimensional step function with jumps p_i at $\boldsymbol{t}_1, ..., \boldsymbol{t}_n$,

$$p_i = G(t_{ix}, t_{iy}) - G(t_{i-1,x}, t_{iy}) - G(t_{ix}, t_{i-1,y}) + G(t_{i-1,x}, t_{i-1,y}), \quad i = 1, ..., n.$$

Defining $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)'$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_m')'$, then the empirical likelihood is given by,

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, p_1, ..., p_n) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} w_1(\boldsymbol{x}_{1j}) \cdots \prod_{j=1}^{n_m} w_m(\boldsymbol{x}_{mj})$$

where $w_j(\boldsymbol{t}) = \exp\{\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(\boldsymbol{t})\}$. The likelihood is maximized with respect to the parameters subject to the constraints

$$\sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i[w_1(\boldsymbol{t}_i) - 1] = 0, \ldots, \ \sum_{i=1}^{n} p_i[w_m(\boldsymbol{t}_i) - 1] = 0$$

in two steps. First, for fixed $\boldsymbol{\alpha}$'s and $\boldsymbol{\beta}$'s, the likelihood is maximized with respect to the p_i to yield

$$p_i \equiv p_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{n_0} \cdot \frac{1}{1 + \rho_1 w_1(\boldsymbol{t}_i) + \cdots + \rho_m w_m(\boldsymbol{t}_i)} \tag{2.8}$$

where the ρ_j are relative sample sizes with respect to n_0,

$$\rho_j = n_j / n_0, \quad j = 1, ..., m.$$

Hence, the optimal p_i are functions of the $\boldsymbol{\alpha}$'s and $\boldsymbol{\beta}$'s. Substituting the p_i back into the likelihood gives a function of the $\boldsymbol{\alpha}$'s and $\boldsymbol{\beta}$'s only, from which we obtain maximum likelihood estimates denoted by $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$. Therefore,

$$\hat{p}_i = p_i(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}),$$

and the estimate of the reference cdf G from the entire fused data $\boldsymbol{t}$ is

$$\hat{G}(\boldsymbol{x}) = \sum_{i=1}^{n} \hat{p}_i I_{(-\infty, \boldsymbol{x}]}(\boldsymbol{t}_i)$$

where $(-\infty, \boldsymbol{x}] = (-\infty, x] \times (-\infty, y]$ for $\boldsymbol{x} = (x, y)$. Note that $I_A(\omega) = 1$ for $\omega \in A$ and $I_A(\omega) = 0$ otherwise. The equidistribution hypothesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = ... = \boldsymbol{\beta}_m = \mathbf{0}$ can be tested by means of the likelihood ratio (LR),

$$LR \equiv -2[\ell(\mathbf{0}, \mathbf{0}) - \ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})]$$

Under H_0, the likelihood ratio is approximately distributed as χ^2 with pm degrees of freedom, and H_0 is rejected for large values.

2.9 Diagnostic Plots

Graphical evidence of goodness-of-fit can be obtained from the plots of $\hat{G}_i$ versus the corresponding empirical multivariate distribution function $\tilde{G}_i, i = 0, 1, ..., m$, evaluated at some selected two-dimensional points as to obtain two dimensional plots. Figures 2.7 - 2.9 in the next section are examples of this. We refer to these plots as diagnostic plots. See Voulgaraki et al. (2012) [34].

2.10 Simulation

In this section, we simulate three cases of the bivariate normal distributions with either equal covariance matrices or unequal covariance matrices. In the present simulation study, $m = 1$ and g_0 denotes the reference distribution.

1. $g_0 \sim N((0,0)', \boldsymbol{\Sigma})$, $g_1 \sim N((0,0)', \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$, $n_0 = 100$, $n_1 = 80$.

2. $g_0 \sim N((1,4)', \mathbf{\Sigma})$, $g_1 \sim N((0,0)', \mathbf{\Sigma})$ with $\mathbf{\Sigma} = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$, $n_1 = 100$, $n_1 = 80$.

3. $g_0 \sim N((0,0)', \mathbf{\Sigma}_0)$, $g_1 \sim N((0,0)', \mathbf{\Sigma}_1)$ with $\mathbf{\Sigma}_0 = \begin{pmatrix} 6 & 1 \\ 1 & 10 \end{pmatrix}$, $\mathbf{\Sigma}_1 = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$, $n_0 = 100$, $n_1 = 80$.

The bivariate normal distribution with the same covariance matrices follows the density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$, but this is not true for the bivariate normal distribution with unequal covariance matrices. The bivariate normal distribution with unequal covariance matrices follows the density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$. Hence we expect to see straight lines in the diagnostic plots in cases (1) and (2) for $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$. On the other hand, we expect to see deviations from straight lines in the diagnostic plots in case (3) for $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$. We should see straight lines in the diagnostic plots in cases (1),(2) and (3) for $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$.

Figures 2.7 - 2.9 show the estimated $\hat{G}_1$ and $\hat{G}_2$ versus the empirical cdf $\tilde{G}_1$ and $\tilde{G}_2$, respectively, all obtained from one run of the simulated data and evaluated at selected points in $\mathbf{R^2}$. As expected, in cases (1) and (2), there is almost a perfect agreement between $\hat{G}_i$ versus $\tilde{G}_i$, $i = 1, 2$ with $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$, whereas the density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ is not appropriate for the data from case (3). The density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$ works well for all cases (1), (2) and (3).

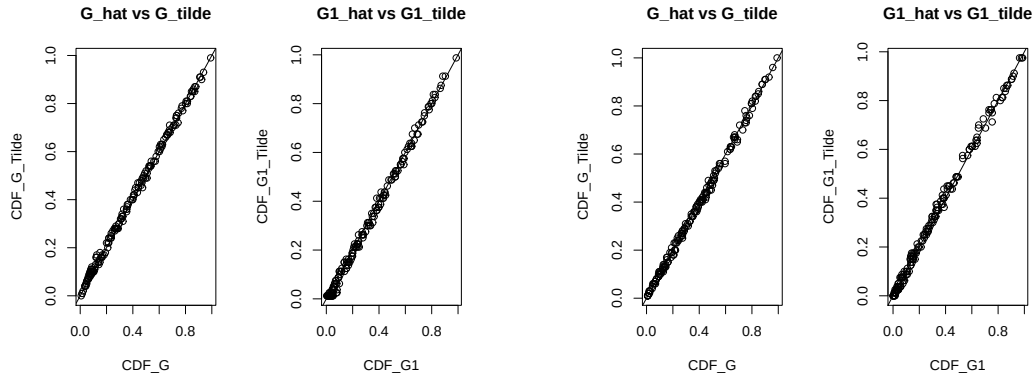

Figure 2.7: Case-control plots of $\hat{G}_i$ vs. $\tilde{G}_i,\ i = 0, 1$, simulations (1) for $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ and $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$.

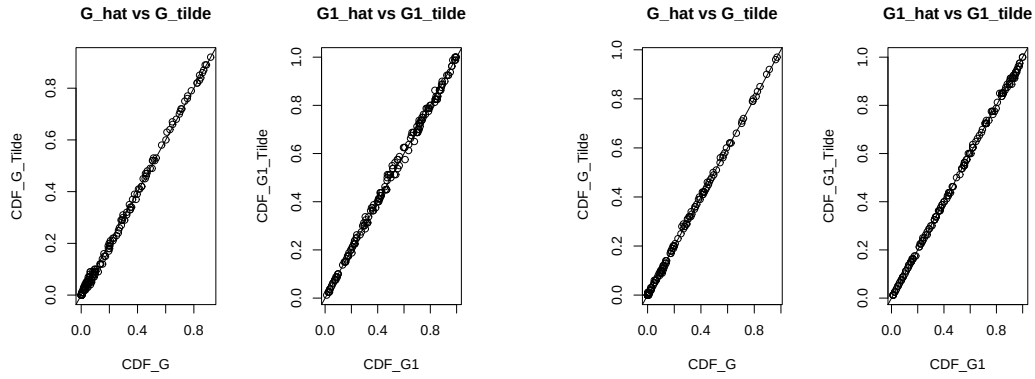

Figure 2.8: Case-control plots of $\hat{G}_i$ vs. $\tilde{G}_i,\ i = 0, 1$, simulations (2) for $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ and $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$.

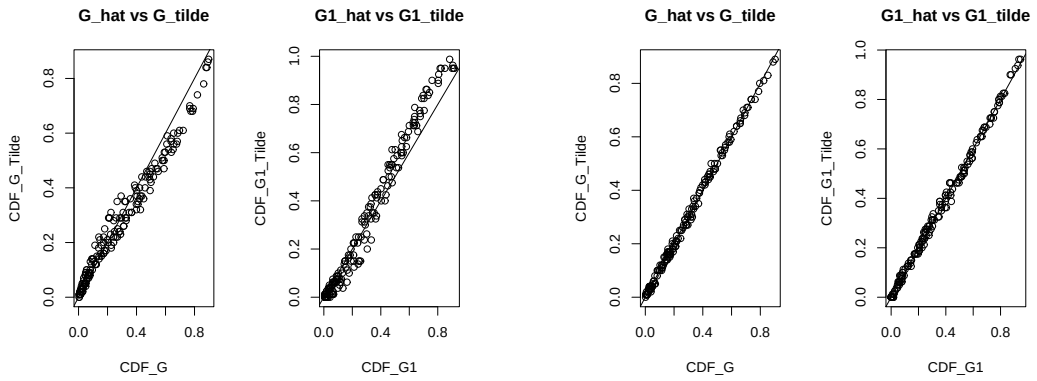

Figure 2.9: Case-control plots of $\hat{G}_i$ vs. $\tilde{G}_i$, $i = 0, 1$, simulations (3) for $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ and $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$.

2.11 Application to Ball Bearing Testing

The Ball Bearing data have strong autocorrelation patterns; see Figure 2.10. We shall apply the bivariate paradigm discussed above. Such a bivariate analysis could address dependence of the ball bearing data. It should give a better performance when deciding good vs faulty.

In practice, we sample pairs of data from (x_t, x_{t-1}) where each sample is of size 100. We have seen for the density ratio model with $h(x) = x$ that when each sample is of size 1000 works well. Here we decrease the sample size to 100. In this case, the density ratio model with $h(x) = x$ may not work well anymore. However, we find that the bivariate density ratio model works well. This indicates higher power for the bivariate density ratio model when discriminating between normal and faulty bearings.

The histograms of the ball bearing data are approximately normal as can be

seen in Figure 2.11. This leads us to use the density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ and $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$. As Figure 2.12-2.13 show, the density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$ is a suitable model for the Ball Bearing data. There is almost perfect agreement between the plots of the semiparametric $\hat{G}_i$ and the corresponding empirical $\tilde{G}_i, i = 1, 2$. However the density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ does not preform as well as the model with the quadratic $\boldsymbol{h}$.

We apply the density ratio model with $h(x) = x$, $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ and $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$ in five different trials. Table 2.1 shows the p-values from the likelihood ratio test for BadX275 versus GoodX098. The density ratio models with $h(x) = x$ or $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ sometimes fail to discriminate between normal and faulty bearings. However, the density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$ works very well. Table 2.2 shows the p-values from the likelihood ratio test for BadX279 versus GoodX098. We observe similar situation as in Table 2.1.

One the other hand, when both samples are from the normal vibration GoodX098, the p-values jump dramatically upward as they should. See Table 2.3.

We see that, by bringing in dependence of the ball bearing data, we obtain better performance when deciding good vs faulty, which points to the potential of the semiparametric method in effective discrimination between normal and faulty bearings when the sample sizes are not large enough.

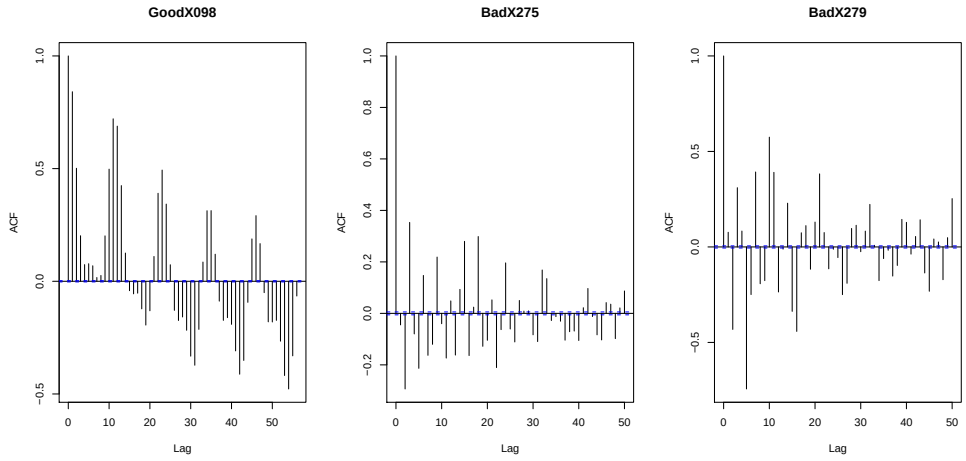

Figure 2.10: ACF plots corresponding to GoodX098, BadX275, BadX279.

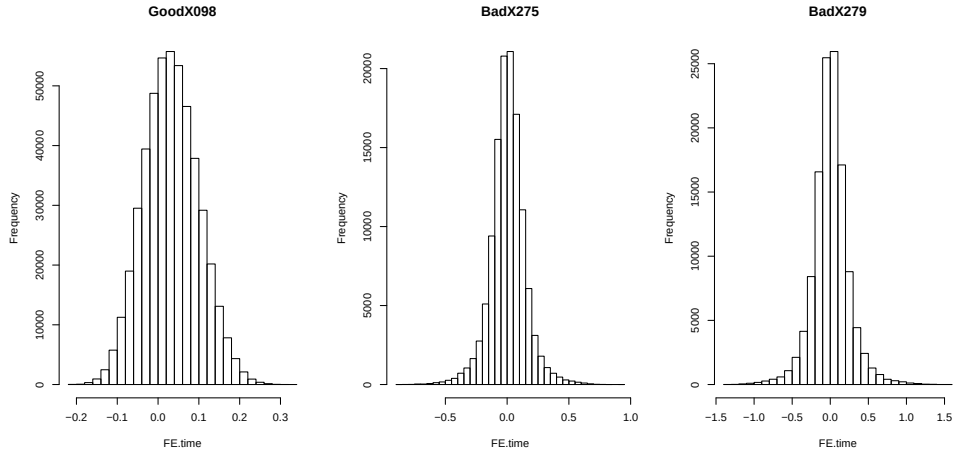

Figure 2.11: Histograms corresponding to GoodX098, BadX275, BadX279.

Table 2.1: BadX275 versus GoodX098

p-values	Tilt Function	1	2	3	4	5
Univariate	$h(x) = x$	0.30749	0.008468	0.20958	0.59889	0.25928
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$	0.43202	0.007888	0.02413	0.08903	0.04803
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$	0	0	0	0	0

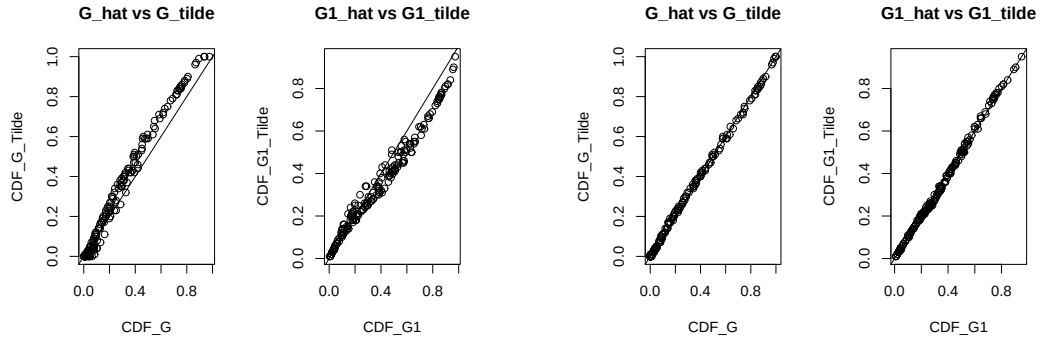

Figure 2.12: Case-control plots of $\hat{G}_i$ vs. $\tilde{G}_i,\ i = 1, 2$, BadX275 versus GoodX098 for $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ and $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$.

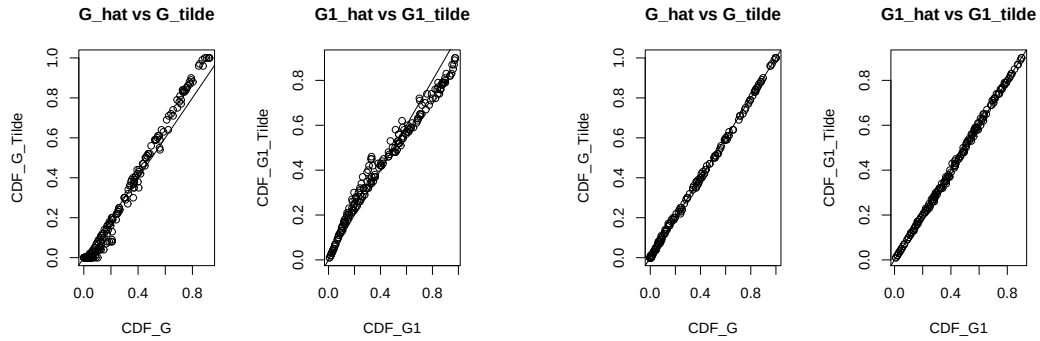

Figure 2.13: Case-control plots of $\hat{G}_i$ vs. $\tilde{G}_i,\ i = 1, 2$, BadX279 versus GoodX098 for $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ and $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$.

Table 2.2: BadX279 versus GoodX098

p-values	Tilt Function	1	2	3	4	5
Univariate	$h(x) = x$	0.004542	0.65365	0.01377	0.26823	0.39945
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$	0.005668	0.10939	0.04706	0.50404	0.33879
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$	0	0	0	0	0

Table 2.3: GoodX098 versus GoodX098

p-values	Tilt Function	1	2	3	4	5
Univariate	$h(x) = x$	0.91614	0.92374	0.54844	0.70672	0.54905
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$	0.73661	0.46588	0.82990	0.77055	0.60365
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$	0.95975	0.53210	0.45631	0.97805	0.69764

2.12 Application to Motor Testing

The Motor data have strong autocorrelation patterns. See Figure 2.14. We apply the Bivariate paradigm to the Motor data. Such a bivariate analysis could bring in dependence of the Motor data. It should give better performance when deciding good vs faulty.

In practice, we sample pairs of data from (x_t, x_{t-1}) where each sample is of size 100. We have seen for the density ratio model with $h(x) = x$, and each sample is of size 500, the DRM works well for Motor data. Here we decrease the sample size to 100, in this case, the density ratio model with $h(x) = x$ may not work well anymore, however, we find that the bivariate density ratio model works well. This indicates higher power for the bivariate density ratio model when discriminating between normal and faulty motors.

As Figures 2.15-2.16 show, the density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$ is a suitable model for the Motor data. There is almost a perfect agreement between the plots of the semiparametric $\hat{G}_i$ and the corresponding empirical $\tilde{G}_i, i = 1, 2, 3$. The density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ also works pretty well, but does not perform as well as the previous one.

We apply the density ratio model with $h(x) = x$, $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ and $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$ in five different trials. Table 2.4 shows the p-values from the likelihood ratio test for three healthy signatures. The large p-values indicate the hypothesis H_0 of equidistribution is accepted quite convincingly in each case.

On the other hand, when a single "bad" signature replaces a healthy one, the p-values were reduced dramatically as they should; see Table 2.5. Density ratio model with $h(x) = x$, $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ sometimes fail to discriminate between normal and faulty motors. However, density ratio model with $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$ works very well.

When two "bad" signatures replaced healthy signatures, the p-values were again very small. Table 2.6 shows the p-values from the likelihood ratio test for two "bad" signatures. We observe similar situation as in Table 2.5.

We see that, by bringing in dependence of the motor data, we obtain better performance when deciding good vs faulty, which points to the potential of the semiparametric method in effective discrimination between normal and faulty motors when sample size is not large enough.

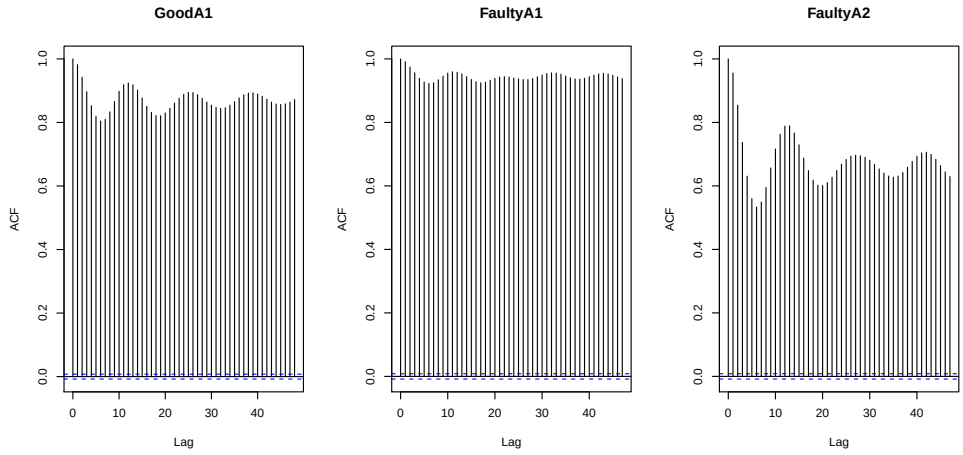

Figure 2.14: ACF plots corresponding to GoodA1, FaultyA1, FaultyA2.

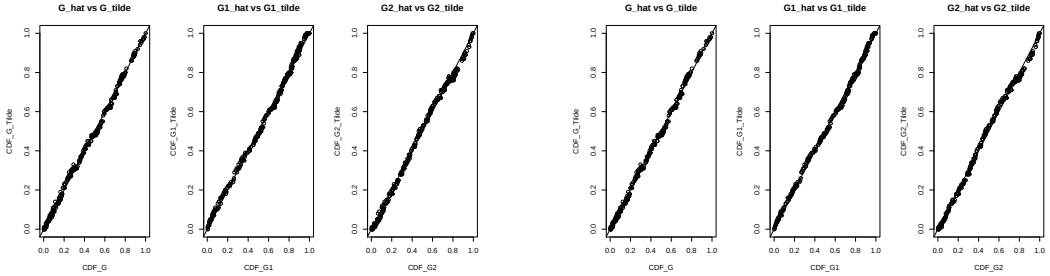

Figure 2.15: Case-control plots of $\hat{G}_i$ vs. $\tilde{G}_i$, $i = 1, 2, 3$, A Single "Bad" Signature for $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ and $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$.

Table 2.4: Three Healthy Signatures

p-values	Tilt Function	1	2	3	4	5
Univariate	$h(x) = x$	0.92311	0.78235	0.84094	0.54022	0.58041
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$	0.42933	0.29146	0.76423	0.52217	0.47945
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$	0.30640	0.47171	0.40879	0.12281	0.88390

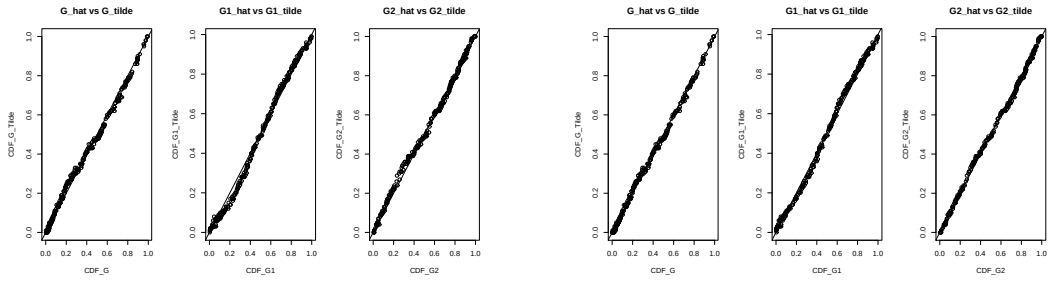

Figure 2.16: Case-control plots of $\hat{G}_i$ vs. $\tilde{G}_i,\ i = 1, 2, 3$, Two "Bad" Signatures for $\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$ and $\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$.

Table 2.5: A Single "Bad" Signature

p-values	Tilt Function	1	2	3	4	5
Univariate	$h(x) = x$	0.003426	0.301305	0.000900	0.006887	0.000131
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$	0.000761	0.181557	0.000728	0.006724	6.61e-05
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$	1.34e-08	6.62e-07	5.85e-08	3.15e-06	1.08e-09

Table 2.6: Two "Bad" Signatures

p-values	Tilt Function	1	2	3	4	5
Univariate	$h(x) = x$	0.006587	0.196054	0.000985	0.003143	0.000137
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x, y)'$	0.004379	0.147094	0.000893	0.002989	0.000105
Bivariate	$\boldsymbol{h}(\boldsymbol{x}) = (x^2, x, y^2, y, xy)'$	4.13e-13	7.04e-11	3.95e-13	8.21e-11	6.88e-12

2.13 Discussion

We have illustrated a semiparametric approach to monitoring the operating conditions of vibrating machines based on the fusion or integration of several samples from different acceleration signals. It has been shown that the semiparametric fusion method was quite effective in testing equidistribution, giving high p-values when the hypotheses were accepted, and extremely low p-values when the hypotheses were rejected. Hypothesis tests actually defines a distance between distributions, which is potentially useful in clustering analysis based on distributions.

Chapter 3: Small Area Estimation

3.1 Introduction

Sample surveys are widely used to estimate population attributes, such as totals, means and other parameters of finite populations by obtaining data from a subset of a population. In many applications, the same information is also desired for small sub-populations such as individuals in small geographical area such as a county or particular demographic within an area. Often, the surveys are carried out for the population as a whole (for example, a nation or similarly high levels). The sample size within many subpopulations of interest can be very small or even without sampling units due to the random nature of probability sampling. Estimating these subpopulations attributes with good accuracy is a interesting problem. To deal with this problem, it could use additional data (such as census records) that exist for these small areas in order to obtain estimates. Most existing methods have focused on estimating small area means. There are fewer discussion in estimating small are quantiles. In this Chapter, we assume the small area population distributions have a linear structure with error terms satisfying the density ratio model (DRM). That is, the small area error distributions are all tilted distributions relative to a common base distribution. We choose the basis or reference distribution as the distribution

of computer generated data provided the DRM passes a goodness of fit test. That is, we fuse the real data and artificial data. This approach allows us to give a distribution estimation of the small area population. And by fusing appropriate artificial data, we can increase the sample size (making the small area into a big area) and control the common base distribution to get better estimates. We also suggest estimators for population attributes of subpopulations with no sampling units.

3.2 Small Area Estimation

We consider the nested-error unit level regression model (NER) of Battese et al. (1988) [1]. The whole population consists of $m+1$ small areas and n_k sampling units are obtained from the k-th area ($k = 0, 1, 2, ..., m$). Under this model, the response variable y and the covariates satisfy the following formula:

$$y_{kj} = \boldsymbol{x}_{kj}^{\tau}\boldsymbol{\beta} + v_k + \epsilon_{kj}, \quad j = 1, 2, \ldots, n_k \tag{3.1}$$

where $v_k \sim N(0, \sigma_v^2)$ is the random effects in each small area and $\epsilon_{kj} \sim N(0, \sigma^2)$ is the random errors for each observation. Under this model, the regression coefficient $\boldsymbol{\beta}$ remains the same. Therefore, samples from all areas can be used to estimate $\boldsymbol{\beta}$. Hence, when the whole population sample size $\sum n_k$ is increasing, an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be obtained with any precision. Suppose the area totals $\boldsymbol{X}_k$ are known, then, a direct estimate of the k-th area total would be $\hat{Y}_k = \boldsymbol{X}_k^{\tau}\hat{\boldsymbol{\beta}}$. Various estimation strategies have been proposed based on model (3.1). Most existing methods focus on estimating small area totals or means. There are fewer discussions about small

area medians or quantiles. Some discussions are based on quantile regression. A general reference about quantile regression is [25]. In this Chapter, we construct median and quantile estimators by using the density ratio model.

3.3 Density Ratio Models in Small Area Estimation

Assume that we have a random sample from the target finite population with n_k units from the k-th small area, and that there are m small areas in the population. Chen and Liu (2015) [6] suggest that

$$y_{kj} = \boldsymbol{x}_{kj}^{\tau}\boldsymbol{\gamma}_k + \epsilon_{kj} \tag{3.2}$$

Some specifications of this model are as follows. First, we allow a more flexible linear relationship with area specific regression coefficient $\boldsymbol{\gamma}_k$, but forgo the area specific random effect in model (3.1). To avoid excessive numbers of parameters in this model, we need to seek a way to link $\boldsymbol{\gamma}_k$ to some auxiliary information. There are many potential choices, Chen and Liu (2015) [6] suggest that

$$\boldsymbol{\gamma}_k = \boldsymbol{\gamma} + a\bar{\boldsymbol{X}}_k \tag{3.3}$$

for some vector $\boldsymbol{\gamma}$ and scalar a, where the $\bar{\boldsymbol{X}}_k$'s are known area specific means of covariates. Second, regard ϵ_{kj}, for each k, as a random sample from some distribution G_k. Chen and Liu (2015) [6] postulate a density ratio model such that for $k = 1, 2, ..., m$

$$\log\{dG_k(t)/dG_0(t)\} = \alpha_k + \boldsymbol{\beta}_k'\boldsymbol{h}(t) \tag{3.4}$$

This idea occurred first in Kedem et al.(2008) [20]. We choose the base distribution G_0 by computer generated data provided model (3.4) passes a goodness of fit test. Observe that we fuse real and artificial data. By fusing the real data with appropriate artificial data, we can increase the sample size (making the small area into a big area) and control the common base distribution G_0 to get better estimates of the DRM.

Given $(y_{kj}, \boldsymbol{x}_{kj})$ for $k = 1, ..., m$ and $j = 1, ..., n_k$, we can estimate $(\boldsymbol{\gamma}, a)$ in (3.3) through least squares. That is, let

$$(\hat{\boldsymbol{\gamma}}, \hat{a}) = \underset{\boldsymbol{\gamma}, a}{\operatorname{argmin}} \sum_{k,j} \{y_{kj} - \boldsymbol{x}_{kj}^{\tau}(\boldsymbol{\gamma} + a\bar{\boldsymbol{X}}_k)\}^2. \tag{3.5}$$

Hence, the residuals $(k = 1, ..., m; j = 1, 2, ..., n_k)$ are given by

$$\hat{\epsilon}_{kj} = y_{kj} - \boldsymbol{x}_{kj}^{\tau}(\hat{\boldsymbol{\gamma}} + \hat{a}\bar{\boldsymbol{X}}_k). \tag{3.6}$$

Then we can follow the estimation method which was discussed in Section 1.5 by using $\hat{\epsilon}_{kj}$. The corresponding estimated distorted cdfs of the residuals are given by,

$$\hat{G}_k(\epsilon) = \sum_{i=1}^{n} \exp(\hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k' \boldsymbol{h}(t_i))\hat{p}_i I(t_i \le \epsilon) \quad k = 1, 2, \ldots, m, \tag{3.7}$$

where $\mathbf{t} = (\mathrm{t}_1, ..., \mathrm{t}_n)' = (\boldsymbol{\epsilon}_0', \hat{\boldsymbol{\epsilon}}_1', ..., \hat{\boldsymbol{\epsilon}}_m')'$, $\boldsymbol{\epsilon}_0$ is a vector of computer generated data provided model (3.4) passes a goodness of fit test.

The availability of $\hat{G}_k$ provides a new tool for small area estimation. If the small area mean $\bar{y}_k$ is the parameter of interest, we can estimate it by

$$\hat{\bar{y}}_k = \bar{\boldsymbol{X}}_k^{\tau}\hat{\boldsymbol{\gamma}}_k + \int \epsilon d\hat{G}_k(\epsilon) \tag{3.8}$$

$$= \bar{\boldsymbol{X}}_k^{\tau}(\hat{\boldsymbol{\gamma}} + \hat{a}\bar{\boldsymbol{X}}_k) + \int \epsilon d\hat{G}_k(\epsilon) \tag{3.9}$$

Following Kedem et al. (2008) [20], the small area distribution of y can be comprehensively estimated as

$$\begin{aligned}
\hat{F}_k(y) &= \hat{P}(y_k \leq y) && (3.10)\\
&= \hat{P}(\boldsymbol{x}_k^\tau \boldsymbol{\gamma}_k + \epsilon_k \leq y) && (3.11)\\
&= \hat{P}(\epsilon_k \leq y - \boldsymbol{x}_k^\tau \boldsymbol{\gamma}_k) && (3.12)\\
&= \hat{G}_k(y - \boldsymbol{x}_k^\tau \boldsymbol{\gamma}_k) && (3.13)\\
&\approx \hat{G}_k(y - \bar{\boldsymbol{X}}_k^\tau \{\hat{\boldsymbol{\gamma}} + \hat{a}\bar{\boldsymbol{X}}_k\}) && (3.14)
\end{aligned}$$

We may hence estimate the small area quantiles by those of $\hat{F}_k(y)$.

3.4 Dealing with missing data

The random nature of the probability sampling can result in no sampling units from many subpopulations of interest. From our original model (3.2), we are supposed to know y_{kj} and $\boldsymbol{x}_{kj}$. In reality, sometimes we don't know the values of y_{kj} or of $\boldsymbol{x}_{kj}$. Without loss of generality, we assume we have missing values in the m-th area.

3.4.1 Missing covariates

If the covariates $\boldsymbol{x}_{kj}$ are missing, we can apply model (3.2) to the rest of the $m-1$ areas, and get $\hat{G}_k$ for $k = 1, 2, ..., m-1$. We suggest to estimate the small

area distribution of y in the m-th area by the average

$$\hat{F}_m(y) = (m-1)^{-1} \sum_{k=1}^{m-1} \hat{G}_k(y - \bar{Y}_m) \tag{3.15}$$

Then we may estimate the small area mean or quantiles for the m-th area from $\hat{F}_m(y)$.

3.4.2 Missing variable of interest

If the variable of interest y_{kj} is missing, we can apply model (3.2) to the rest of the $m-1$ areas, and get $\hat{\boldsymbol{\gamma}}_k$. Then we can create the estimated $\hat{y}_{mj} = \boldsymbol{x}_{mj}^{\tau} \hat{\boldsymbol{\gamma}}_k$. We suggest estimating the small area distribution of y in the m-th area by

$$\hat{F}_m(y) = (m-1)^{-1} \sum_{k=1}^{m-1} \hat{G}_k(y - \bar{\hat{Y}}_m) \tag{3.16}$$

The small area mean or quantiles for the m-th area can be estimated from $\hat{F}_m(y)$.

3.5 Simulation

In this section, we do several numerical simulations to investigate the performance of the estimator (3.14) for small area quantiles. We take the 10%, 25%, 50%, 75% and 90% small area quantiles as the parameters of interest. We simulated data from three models where the number of small areas is $m = 10$. In each case, the small area sample size is $n_k = 2$, 10, 50 or 100. The tilt function $\boldsymbol{h}(t)$ in (3.4) is set to t. We use several GOF tests to verify $\boldsymbol{h}(t) = t$ is an appropriate tilt function. The process is repeated independently $N = 100$ times. Let Y_k denote the theoretical small area quantiles for the k-th area, and let $\hat{Y}_k^i$ denote the estimated small area

quantiles for the k-th area in the i-th repetition. We report the average mean square error (AMSE) defined as

$$AMSE = (Nm)^{-1} \sum_{k=1}^{m} \sum_{i=1}^{N} (\hat{Y}_k^i - Y_k)^2$$

next, we specify three models used in this simulation. In these models, the covariates $\boldsymbol{x}$ and response value y are linked as follows,

$$y_{kj} = \boldsymbol{x}_{kj}^{\tau}\boldsymbol{\beta} + \epsilon_{kj} \tag{3.17}$$

$$y_{kj} = \boldsymbol{x}_{kj}^{\tau}(\boldsymbol{\beta} + \bar{\boldsymbol{X}}_k/2) + \epsilon_{kj} = \boldsymbol{x}_{kj}^{\tau}\boldsymbol{\beta} + \boldsymbol{x}_{kj}^{\tau}\bar{\boldsymbol{X}}_k/2 + \epsilon_{kj} \tag{3.18}$$

$$y_{kj} = \boldsymbol{x}_{kj}^{\tau}(\boldsymbol{\beta} + \bar{\boldsymbol{X}}_k/2) + \mu_k + \epsilon_{kj} = \boldsymbol{x}_{kj}^{\tau}\boldsymbol{\beta} + \boldsymbol{x}_{kj}^{\tau}\bar{\boldsymbol{X}}_k/2 + \mu_k + \epsilon_{kj} \tag{3.19}$$

where the first component of $\boldsymbol{x}_{kj}$ is 1, the other two components of $\boldsymbol{x}_{kj}$ are k, $k+1$. $\boldsymbol{\beta} = c(1, 1, -1)$, ϵ_{kj} and μ_k are standard normal.

The simulation results on average mean square errors of the estimators for small area quantiles are presented in Tables 3.1, 3.2 and 3.3. we can observe that as the small area sample size n_k increasing, the average mean square error is decreasing.

Table 3.1: Simulation 1, m=10, N=100, fused with Norm(0,1) with size n_k

AMSE	10%	25%	50%	75%	90%
n_k=2	0.7122751	0.5649076	0.5134292	0.5847592	0.7784737
n_k=10	0.1260628	0.1136636	0.112306	0.1126443	0.1267265
n_k=50	0.02514846	0.02370558	0.02368818	0.02371201	0.02732065
n_k=100	0.01313619	0.01213875	0.01154132	0.01148632	0.01266889

Table 3.2: Simulation 2, m=10, N=100, fused with Norm(0,1) with size n_k

AMSE	10%	25%	50%	75%	90%
n_k=2	0.7682328	0.6058363	0.5576685	0.6234982	0.8191698
n_k=10	0.1236161	0.1130988	0.1127212	0.1141786	0.1260709
n_k=50	0.02397778	0.02317001	0.02258742	0.02315008	0.02611703
n_k=100	0.01347736	0.01198986	0.01172397	0.01225897	0.01309221

Table 3.3: Simulation 3, m=10, N=100, fused with Norm(0,1) with size n_k

AMSE	10%	25%	50%	75%	90%
n_k=2	0.8054222	0.5962571	0.5080348	0.5842774	0.8079982
n_k=10	0.1337086	0.1151482	0.1047073	0.1103459	0.133382
n_k=50	0.02745751	0.02374751	0.02318081	0.02481688	0.02942381
n_k=100	0.01376919	0.01150948	0.01076977	0.0112752	0.01268921

3.6 LANDSAT data

We use the LANDSAT data (Battese et al. 1988 [1]). The initial survey data, in which farmers reported the area they had growing either corn or soybeans, was compared to estimates obtained from satellite mapping of the farms. The landsat data.frame in R is a compilation of survey and satellite data. It consists of data on 36 segments under corn and soybeans for 12 counties in north-central Iowa; some of the counties consist only one segment. A segment is about 250 hectares. We report the quantile estimates for corn and soybeans in 12 Iowa counties in Table 3.4 and 3.5.

Table 3.4: Quantile estimates for Corn in 12 Iowa Counties

	10%	25%	50%	75%	90%	mean
Cerro	153.7719	161.8645	165.6329	171.5009	173.7787	165.7600
Hamilton	85.10395	92.17552	96.32000	101.19181	105.81617	96.3200
Worth	69.83421	69.83421	72.74402	76.08000	86.76514	76.0800
Humboldt	132.8490	143.4972	149.3294	162.2817	163.3996	150.8900
Franklin	147.9525	151.0414	158.2226	159.7092	173.4468	158.6233
Pocahontas	90.35912	98.45173	102.22017	108.08812	110.36598	102.5233
Winnebago	100.3369	103.7099	116.9083	119.6163	126.3191	112.7733
Wright	130.0451	141.9561	143.4922	151.4154	154.3395	144.2967
Webster	106.3251	114.6170	117.5411	122.4129	127.0373	117.5950
Hancock	96.28048	101.68602	113.37087	115.55984	122.26265	109.3820
Kossuth	98.15091	98.26412	110.90732	118.79528	127.80509	110.2520
Hardin	111.2013	112.3192	120.9320	125.9133	131.2021	120.0540

Table 3.5: Quantiles estimates for Soybeans in 12 Iowa Counties

	10%	25%	50%	75%	90%	mean
Cerro	-1.026404	2.550684	6.191728	17.060427	18.694347	8.09000
Hamilton	95.96467	101.24946	105.08107	114.06589	115.68542	106.03000
Worth	85.67646	100.02291	105.30771	110.08824	118.12413	103.60000
Humboldt	22.86738	28.80660	34.78524	44.99331	47.68262	35.14500
Franklin	38.86547	38.86547	53.31749	59.34619	66.41038	52.47333
Pocahontas	105.6021	111.8626	122.1129	124.3886	125.9242	118.69667
Winnebago	79.05056	84.48561	85.87976	92.39021	102.10850	88.57333
Wright	79.53198	86.62884	102.87018	104.94914	114.45423	97.80000
Webster	99.93526	105.82442	116.03249	118.72180	120.25743	112.98000
Hancock	107.2181	109.1863	115.3553	126.2630	134.2117	117.47800
Kossuth	99.5183	115.4981	117.5771	127.0822	133.0074	117.84400
Hardin	93.25324	96.83033	100.47137	111.18308	112.73422	101.83400

3.7 Discussion

The small area estimation of population quantiles is not fully discussed in the literature. The currently used models for small area estimation are not suitable as platforms for addressing this issue. We can use density ratio models for the purpose of small area quantile estimation. We choose the reference distribution as that of computer generated data provided the DRM passes a goodness of fit test. And by fusing real data with appropriate artificial data, we can increase the sample size (making the small area big area) and control the base distribution to get better estimates. We also suggest estimators for population attributes of subpopulations with no sampling units.

Chapter 4: Extreme Value Theory

4.1 Introduction

Often, it is required to estimate the probability of rare and hazardous events in many disciplines, including structural engineering, earth sciences, geological engineering and traffic prediction etc. Extreme Value Theory (EVT) is a branch of statistics that deals with extreme deviations from the median of probability distributions. This chapter briefly reviews the Extreme Value Theory (EVT), two wildly used methods in modeling extreme values will be discussed: the block maxima approach and the peaks over threshold approach. The purpose of this Chapter is to review traditional methods in estimation of tail probabilities which will later serve as benchmarks to assess the performance of semiparametric methods. More rigorous and thorough treatment of EVT can be found in Beirlant et al.(2004) [2], Coles (2001) [7], Haan and Ferreira (2006) [9], Leadbetter et al. (1983) [27], and Resnick (1987) [33]. Section 4.2 provides model formulation. We discuss the Block Maxima approach in Section 4.3, and the peaks over threshold approach is given in Section 4.4. The notations are adopted from Coles (2001) [7].

4.2 Model Formulation

Consider a sequence of independent and identically distribution (i.i.d.) random variables $X_1, ..., X_n$ with common distribution function F. The extreme value model focuses on the statistical behavior of

$$M_n = \max\{X_1, ..., X_n\},$$

which is the maximum of the sequence of random variables. Determining which distribution M_n follows is the essential problem in EVT. Theoretically, the distribution of M_n can be derived exactly, given that the distribution function F of X_i is known:

$$P(M_n \leq z) = P(X_1 \leq z, ..., X_n \leq z) = P(X_1 \leq z)...P(X_n \leq z) = (F(z))^n. \quad (4.1)$$

However, this approach is not useful in practice. First, the distribution function F is unknown in general. One possible approach is estimating F by a kernel density estimate. Another approach is assuming that the X_i's come from a particular distribution. Then the estimated F is raised to the power of n to obtain the distribution function of M_n. Small discrepancies in the estimates of F can lead to substantial discrepancies in F^n. Alternatively, a family of distributions F^n that approximate any unknown F may be found. The Fisher–Tippett–Gnedenko theorem provides an asymptotic result. (Fisher and Tippett (1928) [14], Gnedenko (1948) [16])

Theorem 4.1. *Let X_n be a sequence of i.i.d. random variables. If there exist constants $a_n > 0$, $b_n \in R$ and some non-degenerate distribution function G such that*

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G,$$

then G belongs to one of the three standard extreme value distributions: Gumbel, Fréchet, or Weibull distributions.

This is the first EVT result (also known as the Fisher-Tippett-Gnedenko Theorem [14] [16]) which characterizes the asymptotic distribution of the sample maximum. The theorem states that the asymptotic distribution G of the maximum of a sample of i.i.d. random variables, after proper renormalization, can converge in distribution to only one of three possible types of distributions: *Gumbel*, *Fréchet*, or *Weibull*. The three types of distributions correspond to the different tail behaviors for the distribution of the original population. *Gumbel* is related to light-tailed distributions such as normal, gamma or exponential distributions. *Fréchet* is related to heavy-tailed distributions such as Pareto, Cauchy or Student-distribution and *Weibull* is related to distributions with finite upper bound such as Uniform and Beta. These three classes of distributions are termed as the extreme value distribution (EVD).

A reformulation of Theorem 4.1 combines the three distributions into a single family of models called the generalized extreme value (GEV) distribution.

Theorem 4.2. *Let X_n be a sequence of i.i.d. random variables. If there exist constants $a_n > 0$, $b_n \in R$ and some non-degenerate distribution function G such that*

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G,$$

then G a member of the GEV family:

$$G(z) = \exp\Big\{ -\big[1 + \xi\big(\frac{z-\mu}{\sigma}\big)\big]^{-1/\xi}\Big\} \tag{4.2}$$

defined on $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$ *where* $-\infty < \mu < \infty$ *is the location parameter,* $\sigma > 0$ *the scale and* $\xi \neq 0$ *the shape parameter.*

4.3 Block Maxima

The Block Maxima approach considers the maximum the variable takes in successive observations. More precisely, a sample is divided into sub-samples or blocks first. Then, the largest observation in each block (block maximum) is taken as an extreme data point which will be used for fitting the GEV distribution. There are several practical issues when we apply the block maxima approach in a real situation. In the real data application, it is very common that the sample size is not large enough, so that the estimates of the unknown distribution parameters are not reliable. The point estimate and confidence interval come from the block maxima approach don't make any practical sense in this situation.

4.4 Peaks Over Threshold

The peaks over threshold (POT) method is an alternative approach in EVT. It considers all observations above a certain threshold value as extreme observations. The conditional distribution functions of values of x above the threshold u is denoted as F_u. The we need to estimate this conditional excess distribution function. The second EVT result (Pickands-Balkema-de Haan Theorem [3] [30]) provides a very helpful theoretical results that gives the asymptotic distribution of the conditional excess distribution.

Theorem 4.3. *Let X_n be a sequence of i.i.d. random variables with common distribution function F and let*

$$M_n = max(X_1, ..., X_n).$$

Suppose that F satisfies Theorem 4.1, so that for large n, $\frac{M_n - b_n}{a_n} \xrightarrow{d} G$, where

$$G(z) = \exp\Big\{ -\big[1 + \xi\big(\frac{z-\mu}{\sigma}\big)\big]^{-1/\xi}\Big\}$$

Then, for large enough u, the distribution function F_u of $X - u$, conditional on $X > u$, is approximately

$$H(y) = 1 - \Big(1 + \frac{\xi y}{\tilde{\sigma}}\Big)^{-1/\xi}$$

defined on $\{y : y > 0$ and $(1 + \xi y)/\tilde{\sigma} > 0\}$ where $\tilde{\sigma} = \sigma + \xi(u - \mu)$.

The family of distributions determined by H is called the generalized Pareto distribution (GPD). Theorem 4.3 states that, if the limiting distribution of $max(X_1, ..., X_n)$ approximates the GEV distribution G, then the threshold exceedances could be approximated by the generalized Pareto distribution for sufficiently large threshold u. When we apply the POT approach in a real data application, it is important to choose a proper threshold u. If u is too small, a biased sample is obtained. Observations that is not extreme values would be included in the sample and violate the model assumption. On the other hand, if this value is chosen too large, the sample size would be too small. This will cause improper estimation of the parameters in the unknown distribution.

Chapter 5: Out of Sample Fusion and Repeated Out of Sample Fusion

5.1 Introduction

Often, it is required to estimate the probability that a quantity such as mercury, lead, toxicity level, plutonium, temperature, rainfall, damage, wind speed, wave size, risk, etc., exceeds an unsafe high threshold. The probability in question is then very small. To estimate such a probability, information is needed about large values of the quantity of interest. However, in many cases, the data only contain values below or even far below the designated threshold, let alone exceedingly large values. This chapter briefly reviews the Out of Sample Fusion method (OSF) and Repeated Out of Sample Fusion method (ROSF) (Zhou 2013 [39], Katzoff et al. 2014 [24], L. Pan 2016 [29], Kedem et al. 2016 [23]). It is shown that by repeated fusion of the data with externally generated random data, more information about small tail probabilities is obtained with the aid of certain new statistical functions. This provides short, yet reliable interval estimates based on moderately large samples.

5.2 Out of Sample Fusion in Estimation of Threshold Probabilities

Let X_0 denote an i.i.d. sample from some given population:

$$X_0 = (x_{01}, ..., x_{0n_0})' \sim g(x)$$

The distribution function $G(x)$ of X_0 is assumed to be unknown, and the threshold exceedance probability $p = 1 - G(t)$ for some fixed threshold t is of interest. X_0 is referred to as the reference sample. Let X_j denote a computer generated i.i.d. sample with sample size $n_j, j = 1, ..., m$

$$X_j = (x_{j1}, ..., x_{jn_j})' \sim g_j(x)$$

The computer generated samples X_j will be referred to as the fusion samples. Then under the density ratio model, we have

$$\frac{g_j(x)}{g(x)} = \exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x)), \quad j = 1, ..., m$$

where α_j is a scalar parameter, $\boldsymbol{\beta}_j$ is a $p \times 1$ parameter vector, and $\boldsymbol{h}(x)$ is a known $p \times 1$ vector valued distortion or tilt function. Semiparametric statistical inference about all the parameters and probability distribution of the reference X_0 can be obtained from the combined data from the $m+1$ samples $X_0, X_1, ..., X_m$. The combined data now has the size of $n = n_0 + n_1 + ... + n_m$. Therefore. the reference distribution function G is estimated from the fused data with n observations and not just from the reference sample itself with n_0 observations. The estimated threshold exceedance probability is:

$$\hat{p} = \hat{P}(X_0 > t) = 1 - \hat{G}(t) = 1 - \frac{1}{n_0} \sum_{i=1}^{n} \frac{I(t_i \leq t)}{1 + \rho_1 \hat{\omega}_1(t_i) + ... + \rho_m \hat{\omega}_m(t_i)} \tag{5.1}$$

where $\hat{\omega}_j(x) = \exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}_j' \boldsymbol{h}(x), \quad j = 1, ..., m.$

For a large threshold T, the $100(1-\alpha)\%$ confidence intervals for $p = 1 - G(T)$ can be constructed based on the asymptotic results from Theorem 1.2:

$$\left(1 - \hat{G}(t) - z_{1-\alpha/2}\sqrt{\hat{V}(t)}, 1 - \hat{G}(t) + z_{1-\alpha/2}\sqrt{\hat{V}(t)}\right) \tag{5.2}$$

where $\hat{V}_{(t)}$ denotes the estimated variance of $\hat{G}(t)$ as given in Theorem 1.2.

5.3 Repeated Out of Sample Fusion

Repeated Out of Sample Fusion (ROSF) is an extension of OSF to estimate tail probabilities and their confidence intervals where a given reference sample is fused or combined repeatedly with computer generated data. The implementation of ROSF is given in the following.

We want to estimate a small threshold exceedance probability $p > 0$ for a random sample X_0 from some distribution. We call X_0 the reference sample. A fusion sample X_1 is then generated by the computer and fused together with the reference sample. The point estimate $\hat{p}_1$ and the confidence interval $[0, B_1]$ are then obtained through the semiparametric density ratio model as described in the OSF method. The same reference sample is then fused with another computer generated sample (from the same distribution of the previous artificial sample and independent of it) to obtain another $\hat{p}_2$ and confidence interval $[0, B_2]$ in the same manner as before. This process is repeated n_r times to produce a sequence of point estimates $\hat{p}_i$ and confidence intervals $[0, B_i], i = 1, ..., n_r$. Conditional on X_0, the sequence of upper bounds B_i is independent and identically distributed from some distribution

$F_B(\cdot) = F_B(\cdot|X_0)$. Denote the empirical distribution of B_i's by $\hat{F}_B$. By the Glivenko-Cantelli theorem, $\hat{F}_B$ converges to F_B almost surely uniformly as n_r increases. Since the process may be repeated many times, a very close approximation of F_B can be obtained. In other words, as the number of fusions becomes very large, $\hat{F}_B$ is almost the exact F_B.

The final point estimates of the threshold exceedance probability from the ROSF algorithm is the average of $\hat{p}_i$'s from n_r OSF runs:

$$\hat{p} = \hat{P}(X_0 > t) = \frac{1}{n_r}\sum_{i=1}^{n_r}\hat{p}_i, \quad i = 1, ..., n_r,$$

and the associated $100(1-\alpha)\%$ confidence interval is

$$[0, F_B^{-1}(\alpha^{1/N})],$$

where N is a large enough positive integer. More can be found in Kedem et al. (2017) [21], Zhou (2013) [39], Pan (2016) [29] and Kedem et al. (2016) [23].

The length of the confidence interval depends on the choice of N. Here, N serves as a tuning parameter. Intuitively, as the number of fusions increases, the number of B_i's grows and the confidence interval $[0, \max(B_i)]$ covers p with probability close to one. That is, as $n_r \to \infty$.

$$P(B_{(n_r)} > p) \to 1.$$

In practice, the exact CDF of B's F_B is unknown. So the corresponding empirical distribution $\hat{F}_B$ is estimated based on B_i's obtained from n_r OSF repetitions. As $n_r \to \infty$, $\hat{F}_B \to \hat{F}_B$ uniformly almost surely. Therefore, as we control the number of repetitions n_r, F_B is practically known. In Pan (2016) [29], a comparison of ROSF

with a method from extreme values theory (Peaks over Threshold, or POT) points to the merit of this approach.

Similarly, we can get lower bounds $A_1, A_2, ..., A_{n_r}$ for $p = 1 - G(T)$. Since we can fuse as many times as we wish, we practically know the cdf F_A from the Glivenko-Cantelli Theorem (uniform convergence):

$$\hat{F}_A \to F(A)$$

The associated $100(1-\alpha)\%$ confidence interval is

$$[F_A^{-1}(1-\alpha^{1/M}), 1]$$

where M is a large enough positive integer.

Theorem 5.1. *Let $\hat{p}_i$ and A_i be the sequence of point estimates of the tail probabilities and its lower confidence bounds obtained by ROSF. Let F_A denote the distribution function of the A's. Under the condition*

$$P(A \leq p) = F_A(p) > 0$$

there exists M_0 such that for all $M > M_0$, the confidence interval for the tail probability p, $[F_A^{-1}(1-\alpha^{1/M}), 1]$ gives at least $100(1-\alpha)\%$ coverage.

Proof. For an i.i.d. sample $A_1, ..., A_M$, denote the minimum by $A_{(1)} = \min(A_i)$. It follows that

$$P(A_{(1)} \leq p) = 1 - (1 - F_A(p))^M$$

If $P(A \leq p) = F_A(p) > 0$, then from the above equation, the probability that the minimum lower bound covers the desired tail probability increases as the tuning

parameter M increases. Conditional on the given sample X_0, for all $M > M_0$, we have the following inequality:

$$1 - (1 - F_A(p))^M \geq 1 - \alpha$$

for some M_0 sufficiently large. The inequality can be rewritten by inverting the distribution function:

$$F_A^{-1}(1 - \alpha^{1/M}) \leq p \leq 1$$

The above relationship implies that the interval $[F_A^{-1}(1 - \alpha^{1/M}), 1]$ covers the true tail probability p with at least $100(1 - \alpha)\%$ confidence for sufficiently large M. $\square$

Together with the maximum $B_{(N)}$ from the upper bounds $B_1, ..., B_N$, we have with high confidence:

$$F_A^{-1}(1 - 0.05^{1/M}) \leq p \leq F_B^{-1}(0.05^{1/N})$$

In practice, due to computational difficulties, we find that the lower bounds obtained by ROSF may be less than 0. So there is fewer advantages for using the lower bounds to construct a precise confidence interval.

Chapter 6: Iterative Method

6.1 Introduction

It is shown that by repeated fusion of the data with externally generated random data, more information about small tail probabilities is obtained with the aid of certain new statistical functions. In this Chapter, a small tail probability is identified with a point on a certain monotone curve. The point on the curve is approached by an iterative procedure against the backdrop of repeated fusion of real and "fake" data. In many cases, this brings about surprisingly precise estimates for small tail probabilities, using moderately large samples. A comparison of the approach with a method from extreme values theory (Peaks over Threshold, or POT), using both artificial and real data, points to the merit of the approach. A preliminary version of our work can be found in Kedem et al. (2018) [19]

6.2 Motivation

We wish to estimate a small tail probability p of exceeding a high threshold T from a moderately large random sample $X_1, ..., X_{n_0}$. This is done by fusing or combining the sample *repeatedly* with computer generated uniform samples. The

number of fusions can be as large as we wish. For example 10,000 or 100,000 or 1,000,000 or more fusions. Throughout the Chapter the sample size n_0 is moderately large (100, 120, or 200), and, since in many cases the data only contain values below or even far below the designated threshold, it is assumed that the measurements X_i are all below T.

Fusing a given sample repeatedly with computer generated data is referred to as *repeated out of sample fusion (ROSF)*. Unlike the bootstrap, additional information is sought repeatedly outside the sample.

The large number of fusions results in what is called a *B-curve* defined in Section 6.4. The B-curve is monotonically increasing and it contains a *point* whose ordinate is very close to p with a high probability. In fact, as the number of fusions increases, the ordinate of that point essentially coincides with p. The goal is to "capture" that point.

Estimating p is equivalent to "capturing" the said point on the B-curve, and this Chapter provides an iterative algorithm for doing so. The consequent interval estimates of p are quite precise. A comparison with peaks-over-threshold (POT) from extreme value theory indicates that ROSF can bring about a substantial gain in reliability as well as in precision across a fairly wide range of tail behavior, given moderately large samples $\boldsymbol{X}_0$.

The question then is how to tie or connect the real data and the generated random data to obtain useful reliable interval estimates for small tail probabilities. Connecting the real and artificial data can be approached by means of their respective distributions under the so called *density ratio model* framework. This Chapter

describes ROSF together with an iterative method (IM) in the estimation of small tail probabilities against the backdrop of the density ratio model.

6.3 A Note about Extremes

The estimation of small tail probabilities has been around for a long time, at least since the celebrated work of Fisher and Tippett (1928) [14] on the extremes of random samples. Of the various statistical methods dealing with this estimation problem, the block maxima (BM) and peaks-over-threshold (POT) are two widespread methods discussed, for example, in Beirlant et al. (2004) [2] and more recently in Ferreira and de Haan (2015) [10], among many others.

BM and POT might not be sufficiently reliable when the data sets are not large enough because both approaches entail a reduction in the number of observations. Specifically, by the POT method only observations above a sufficiently high threshold are used, and by the BM method the data are first divided into blocks and only the maximum from each block is used in estimation. Thus, if the data size is not large to begin with, a further data reduction might reduce considerably the reliability of the estimation results.

ROSF is of an entirely different nature in that it is not based on extreme value theory. It is an *augmentation* method where a reference sample is combined many times with *additional data*, albeit "fake" artificial data. Hence, unlike BM and POT there is no loss of observations.

ROSF has been introduced and applied in the estimation of small tail proba-

bilities in connection with food safety in Kedem et al. (2016) [23]. A large number of experiments show that ROSF accommodates a fairly wide range of tail behavior, including that of gamma, lognormal, inverse Gaussian, Pareto, and Weibull, and that of environmental variables whose distributions possess very long tails, including lead intake, mercury, and chlorophenol compounds.

The special case where only a single fusion with artificial data occurs is dubbed in Zhou (2013) [39] and in Katzoff et al. (2014) [24] as *out of sample fusion* (OSF). In connection with importance sampling, Fokianos and Qin (2008) [13] use this idea in estimating the normalizing constant of a parametric probability distribution. Similarly, Fithian and Wager (2015) [15] study heavy-tailed distributions given a relatively small sample, and a much larger background sample from another distribution assuming that the tails of the two distributions are connected by an exponential tilt model.

The relative efficiency of BM and POT has been discussed in Ferreira and de Haan (2015) [10] and references therein. Under certain conditions the two methods are quite similar. In this Chapter we shall compare ROSF together with its IM companion to POT only, using moderately large samples.

6.4 ROSF and the B-Curve

We are in pursuit of a small tail probability p. It is shown how to construct a curve which contains with a high probability a point whose ordinate is p.

Let $\boldsymbol{X}_0$ denote a given *reference* sample $x_1, ..., x_{n_0}$ from some reference dis-

tribution G, and suppose we wish to estimate a small tail probability p of that distribution. The variable $X \sim g$ could represent quantities such as earthquake magnitude, radioactive contamination, claim amounts, financial returns, poverty levels, wealth, temperature, and so on, and the interest is in the tail probability $p = P(X > T)$ for some relatively high threshold T. Further, suppose we have a way to fuse or combine the reference sample $\boldsymbol{X}_0$ with a computer-generated sample $\boldsymbol{X}_1$. Then $\boldsymbol{X}_0$ can be fused again with another independent computer generated sample which we again denote by $\boldsymbol{X}_1$ ($\boldsymbol{X}_1$ is used generically), and so on. All these $\boldsymbol{X}_1$ samples are independent and are generated in an identical manner and all have the same size n_1. We refer to these computer-generated samples as *fusion samples*. Observe that the fused or combined samples all have size $n_0 + n_1$.

Here is how B-curves are constructed. We fuse the given reference sample $\boldsymbol{X}_0$ with a computer-generated fusion sample $\boldsymbol{X}_1$ from g_1 and obtain a confidence interval $[0, B_1]$ for the small tail probability p. Since p is small we take the lower bound to be 0, and compute the upper bound B_1. We fuse the given reference sample $\boldsymbol{X}_0$ again with another artificial fusion sample $\boldsymbol{X}_1$ from g_1 and get in the same manner as before another confidence interval $[0, B_2]$ for p. This process is repeated many times to produce a long sequence of confidence intervals $[0, B_i]$, $i = 1, 2, ..., n$. Conditional on $\boldsymbol{X}_0$, the sequence of upper bounds $B_1, B_2, ..., B_n$ is then an independent and identically distributed sequence of random variables from some distribution F_B. It is assumed that

$$P(B_1 > p) = 1 - F_B(p) > 0. \tag{6.1}$$

Let $B_{(1)}, B_{(2)}, ..., B_{(n)}$ be the sequence of order statistics from smallest to largest. Then, as $n \to \infty$, $B_{(1)}$ decreases and $B_{(n)}$ increases. Hence, as the number of fusions n increases the plot consisting of the pairs

$$(1, B_{(1)}), (2, B_{(2)}), ..., (n, B_{(n)}) \tag{6.2}$$

contains a point whose ordinate is p with probability approaching 1. It follows that as $n \to \infty$, there is a $B_{(j)}$ which essentially coincides with p. The plot of points consisting of the pairs $(j, B_{(j)})$ in (6.2) is referred to as the *B-curve*. Typical B-curves corresponding to the tail probability $p = P(X > T) = 0.001$ for various reference samples $\boldsymbol{X}_0$ from the indicated distributions or data are shown in Figure 6.1. Notice that to get $p = 0.001$, in each case the threshold T must change accordingly, and that in each plot there is a $B_{(j)}$ nearest or closest to $p = 0.001$.

A key fact of the present approach is that since the fusions can be repeated indefinitely we can approximate the distribution of the B upper bounds arbitrarily closely.

Let $\hat{F}_B$ be the empirical distribution obtained from the sequence of upper bounds $B_1, B_2, ..., B_n$. Then from the Glivenko-Cantelli Theorem, $\hat{F}_B$ converges to F_B almost surely uniformly as n increases. Since the fusion process can be repeated as many times as we wish, *our key idea*, F_B is known for all practical purposes. Assume then that F_B was obtained from numerous fusions, for example $10,000$ fusions. Then, under (6.1), from a random sample $B_1, ..., B_K$, the probability that the maximum $B_{(K)}$ exceeds p,

$$P(B_{(K)} > p) = 1 - F_B^K(p) \tag{6.3}$$

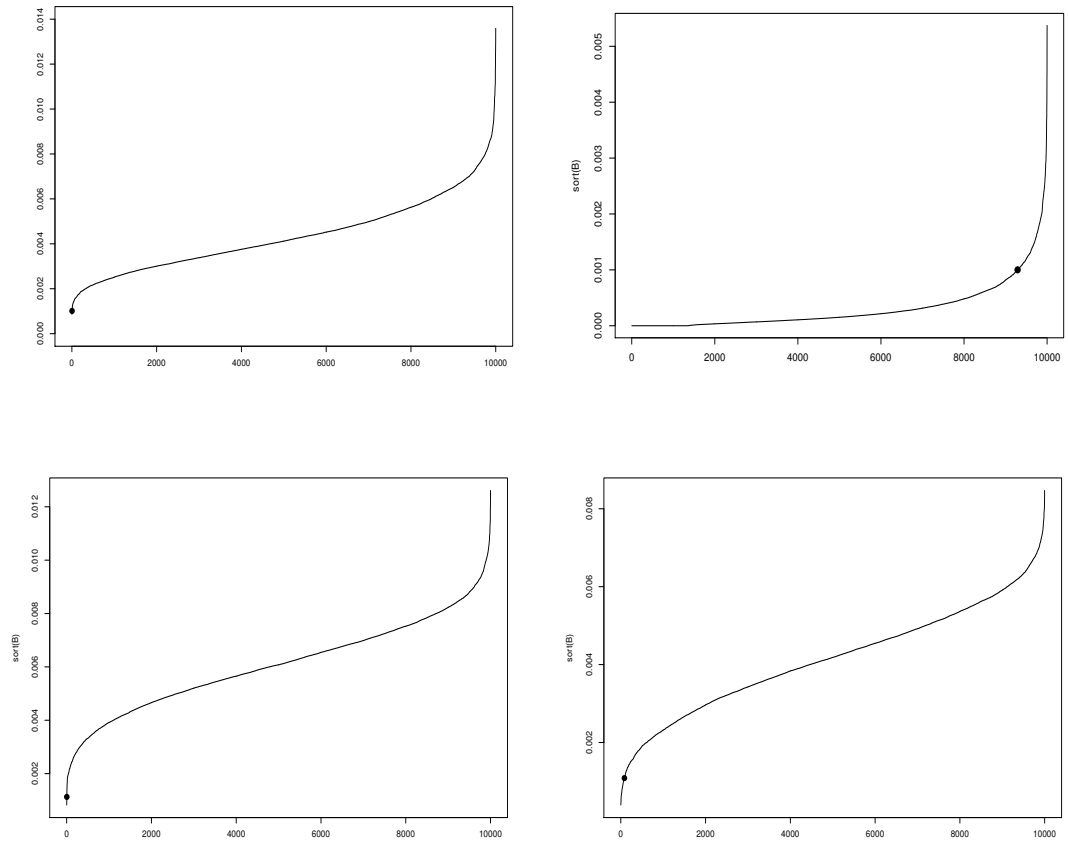

Figure 6.1: Typical B-Curves from $B_{(1)}, ..., B_{(10,000)}$ containing a point corresponding to $p = 0.001$. Clockwise from top left: Gamma(1,0.01), LN(1,1), Lead exposure, Mercury. T=690.7755, 59.7538, 25.00, 22.41, respectively, $n_0 = n_1 = 100$. Histograms representing the distributions are shown in Figure 6.2.

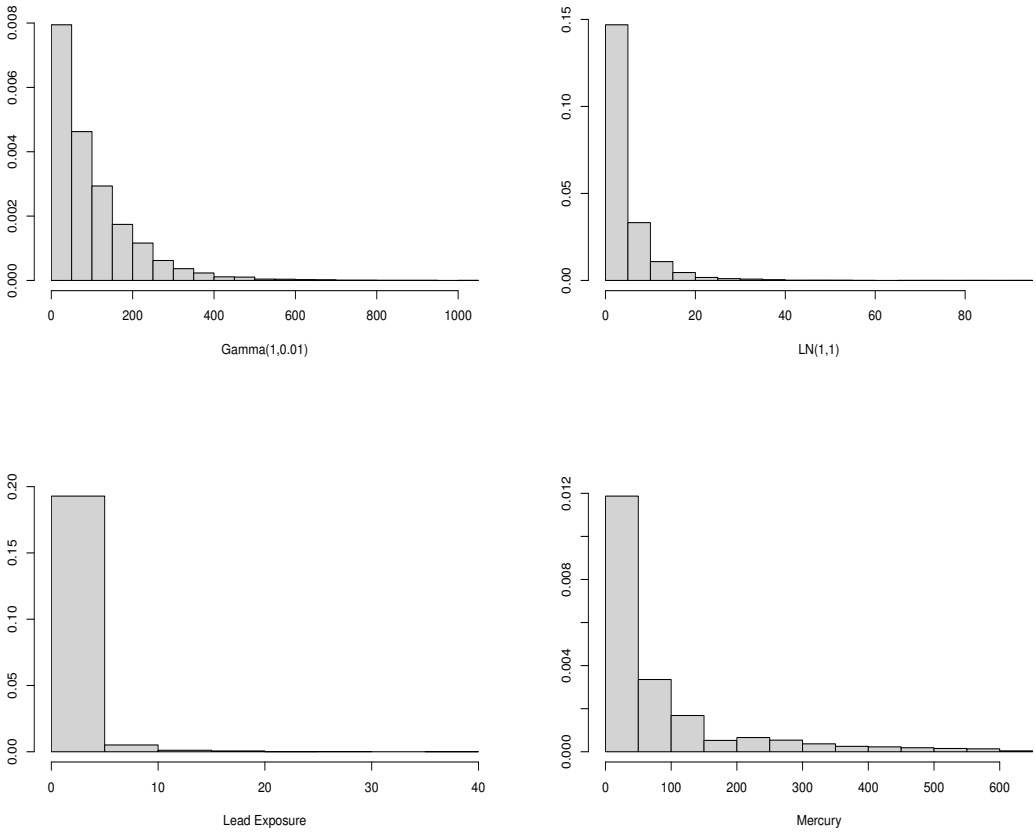

Figure 6.2: Histograms representing distributions with long right tails. The lead intake data are discussed in Kedem et al. (2016) [23]. The mercury data source is NOAA's National Status and Trends Data `https://products.coastalscience.noaa.gov/nsandt_data/data.aspx`

increases with K. It follows that for all $K > K_0$, for some sufficiently large K_0, we have for a small $\alpha > 0$ the inequality

$$1 - F_B^K(p) \geq 1 - \alpha \tag{6.4}$$

or

$$0 < p \leq F_B^{-1}(\alpha^{1/K}). \tag{6.5}$$

The interval (6.5) covers p with at least $100(1-\alpha)\%$ confidence, and it has been applied in food safety in Kedem et al. (2016) [23]. Experimental results indicate that in many cases $K = 100$ is a conservative choice and that at times a much smaller K suffices. However, when $\max(\boldsymbol{X}_0)$ is small relative to T, a larger K is needed, for example $K = 300$ or even larger.

6.4.1 Getting Upper Bounds by Data Fusion

Clearly, the preceding argument is quite general, and the effectiveness of the procedure will depend on the quality of the $[0, B_i]$ confidence intervals. In this section we describe a particular way of generating these confidence intervals, which amounts to *data fusion* of the real and computer-generated data ("augmented reality" as it were) under the density ratio model.

In general, by "fusion" or "data fusion" we mean the combined data from $m + 1$ sources where each source is governed by a probability distribution. In the spirit of augmented reality, random data generating computer algorithms are perfectly legitimate data sources. Using the combined data, semiparametric statistical

inference can be ensued under the density ratio model assumption (Kedem, et al. 2017) [21].

Recall that the reference random sample $\boldsymbol{X}_0$ of size n_0 follows an unknown reference distribution with probability density g, and let G be the corresponding cumulative distribution function (cdf).

Let

$$\boldsymbol{X}_1, ..., \boldsymbol{X}_m,$$

be additional computer-generated random samples where $\boldsymbol{X}_j \sim g_j, G_j$, with size n_j, $j = 1, ..., m$. For now $m \geq 1$ but later we specialize to $m = 1$ only. We refer to the vector

$$\boldsymbol{t} = (t_1, \ldots, t_n)' = (\boldsymbol{X}_0', \boldsymbol{X}_1', \ldots, \boldsymbol{X}_m')', \tag{6.6}$$

of size $n = n_0 + n_1 + \cdots + n_m$ as the fused data. We further assume the *density ratio model* (Qin and Zhang 1997 [31])

$$\frac{g_j(x)}{g(x)} = \exp(\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x)), \quad j = 1, \ldots, m, \tag{6.7}$$

where $\boldsymbol{\beta}_j$ is a $p \times 1$ parameter vector, α_j is a scalar parameter, and $\boldsymbol{h}(x)$ is a $p \times 1$ vector valued distortion or tilt function. Clearly, to generate the $\boldsymbol{X}_j$ samples we must know the corresponding g_j. However, beyond the generating process, we do not make use of this knowledge. Thus, by our estimation procedure, none of the probability densities $g, g_1, ..., g_m$ and the corresponding G_j's, and none of the parameters α's and $\boldsymbol{\beta}$'s are assumed known, but, strictly speaking, the so called tilt function $\boldsymbol{h}$ must be a known function.

Since all the probability distributions are connected by the density ratio model (6.7), each distribution pair g_j, G_j is estimated from the entire fused data $\boldsymbol{t}$ and not just from $\boldsymbol{X}_j$ only. The same holds for the reference pair g, G. Thus, for example, the reference G is estimated from the entire fused data $\boldsymbol{t}$ with n observations and not just from the reference sample $\boldsymbol{X}_0$ with $n_0 << n$ observations.

Under the assumption that the density ratio model (6.7) holds, the maximum likelihood estimate of $G(x)$ based on the fused data $\boldsymbol{t}$ is given in (1.8) in Section 1.5, along with its asymptotic distribution described in Theorem 1.1 and Theorem 1.2. From the theorem we obtain confidence intervals for $p = 1 - G(T)$ for any threshold T using (5.2). In addition, from (5.1) we get the point estimate $\hat{p}$ as well.

Obviously, the density ratio model per se need not hold, and even if it does for some tilt function $\boldsymbol{h}$, the validity or goodness of an arbitrary choice of $\boldsymbol{h}$ is uncertain. Furthermore, if $\max(\boldsymbol{X}_0)$ is much smaller than the threshold T, then $\hat{p}$ from (5.1) could be just too small. However, for the implementation of ROSF, the density ratio model need not hold precisely and any reasonable choice of $\boldsymbol{h}$ suffices, as long as (6.1) holds, which is a mild requirement. Experience shows that the "gamma tilt" $\boldsymbol{h}(x) = (x, \log x)$ is a sensible choice for skewed data similar to those shown in Figure 6.2. Similarly, the "lognormal tilt" $\boldsymbol{h}(x) = (\log x, (\log x)^2)$ is another useful choice.

Our strategy then is to obtain interval estimates for small $p = 1 - G(T)$ for a relatively large T using numerous upper bounds from (5.2), obtained by ROSF, call the upper bounds B_i, and take the lower bounds as 0. This is the method referred to in the previous section by which we obtain the $[0, B_i]$ confidence intervals. When

assumption (6.1) holds, many of the B_i will be greater than p as their number increases, but some will not. Hence, the *ordered* $B_{(i)}$ engulf or surround p with probability approaching one as the number of fusions increase. This is illustrated in Figure 6.1 with 10,000 fusions.

Thus, $[0, B_1]$ is obtained from the first fusion of $\boldsymbol{X}_0$ with a set of m computer-generated samples. Then $[0, B_2]$ is obtained by fusing $\boldsymbol{X}_0$ again but with a *different* independent set of m computer-generated samples, and so on. From each fusion we obtain a point estimate $\hat{p} = 1 - \hat{G}(T)$ using (5.1) and an upper bound B_i from (5.2). Since this fusion process is repeated numerous times, we obtain both numerous point estimates $\hat{p}$'s and numerous upper bounds B_i's. *In general, as the number of fusions increases, the set of pairs* $(j, B_{(j)})$ *engulfs the desired point on the B-curve with probability approaching one.* That is, with a large number of fusions the ordered $B_{(j)}$ engulf p with a high probability. This, in general, cannot be said about the ordered $\hat{p}$'s unless the number of fusions is exceedingly large. See Section 6.5.1.2 for a case where the $\hat{p}$'s from (5.1) are too small.

In this Chapter $m = 1$ only, and the fusion samples are uniform random samples supported over a wide range which covers T. But why uniform? First, when the density ratio model holds for some g and g_1, then it also holds approximately by taking g_1 as a uniform density supported over a sufficiently wide range. Second, and more to the point, ROSF requires only the mild assumption (6.3). Experience shows that assumption (6.3) holds well when fusing $\boldsymbol{X}_0$ with uniform samples using the tilt function $\boldsymbol{h}(x) = (x, \log x)$ *across a wide range of tail types.* Evidently, the B-curves used in this Chapter provide further support for the validity of assumption

(6.3).

To summarize, under assumption (6.3), the B-curves are constructed from ordered upper bounds $B_{(j)}$ (5.2) for $p = P(X > T)$ obtained from a large number of repeated fusions of $\boldsymbol{X}_0$ with random uniform samples $\boldsymbol{X}_1$ where the upper limit of the uniform distribution exceeds T. Throughout the Chapter, $\max(\boldsymbol{X}_0) < T$ and $\boldsymbol{h} = (x, \log x)$.

6.5 Capturing a Point on the B-Curve

Due to a large number of fusions n, F_B is known for all practical purposes and with probability close to 1

$$B_{(1)} < p < B_{(n)}. \tag{6.8}$$

In general, even for $n = 1,000$, $B_{(1,000)}$ is much larger than the true p and $B_{(1)}$ is very close to 0. The goal is to find $B_{(j)}$ close to p.

It follows, by the monotonicity of the B-curve and (6.8), that as j *decreases* (for example from $n = 10,000$), the $B_{(j)}$ approach p from above so that there is a $B_{(j)}$ very close to p. Thus, the B-curve establishes a relationship between j and p.

From a basic fact about order statistics [8] it is known that

$$P(B_{(j)} > p) = \sum_{k=0}^{j-1} \binom{n}{k} [F_B(p)]^k [1 - F_B(p)]^{n-k}. \tag{6.9}$$

Therefore, as (6.9) is monotone decreasing, the *smallest* p which satisfies the inequality

$$\sum_{k=0}^{j-1} \binom{n}{k} [F_B(p)]^k [1 - F_B(p)]^{n-k} \leq 0.95 \tag{6.10}$$

provides another relationship between j and p. Note that if ">" is used instead of "$\leq$" in (6.10) then the solution of (6.10) is $p = 0$. This is so since (6.9) is a steep monotone decreasing step function of the type shown in Figures 6.3, 6.4. Replacing 0.95 by 0.99 in (6.10) gives similar results.

Iterating between these two monotone relationships is what was referred to earlier as the iterative method (IM). The iterative method provides our p estimates. In general, the iteration process could start with a sufficiently large j suggested by the B-curve. With that $j \equiv j_1$ we look for the smallest $p \equiv p_{j_1}$ satisfying (6.10). Next we find a $B_{(j_2)}$ on the B-curve closest to p_{j_1}. This gives a new $j \equiv j_2$ and the previous steps are repeated until convergence occurs and we keep getting the same p. This is our point estimate from the iteration process and it is different than $\hat{p}$ obtained from (5.1).

In symbols, with $B_{(j_k)}$'s from the B-curve, and $p_{(j_k)}$'s the smallest p's satisfying (6.10) with $j = j_k$, and $B_{(j_{k+1})}$ closest to $p_{(j_k)}$, $k = 1, 2, ...,$

$$B_{(j_1)} \to p_{(j_1)} \to B_{(j_2)} \to \cdots B_{(j_k)} \to p_{j_k} \to B_{(j_{k+1})} \to p_{j_k} \to B_{(j_{k+1})} \to p_{j_k} \cdots$$

so that p_{j_k} keeps giving the same $B_{(j_{k+1})}$ (and hence the same j_{k+1}) and vice versa. This can be expressed more succinctly as,

$$j_1 \to p_{(j_1)} \to j_2 \to p_{(j_2)} \to \cdots j_k \to p_{j_k} \to j_{k+1} \to p_{j_k} \to j_{k+1} \to p_{j_k} \cdots$$

As will be illustrated in Section 6.5.1, under some computational conditions this iterative process results in a contraction in a neighborhood of the true p. In a small neighborhood of the true p the $B_{(j_k)}$ can move either up or down, an example of which is given in the lead example in Section 6.5.1.4.

Computationally, the iteration process depends on n and the increments of p at which (6.10) is evaluated. In practice, due to computational limitations of large binomial coefficients the iteration is done as follows. After F_B is obtained from a large number of fusions, 1000 $B_{(j)}$'s are sampled from, say, $n = 10,000$ $B_{(j)}$'s to obtain an approximate B-curve. Next, the binomial coefficients $\binom{n}{k}$ are replaced by $\binom{1000}{k}$. We then iterate between an approximate B-curve and approximate (6.10) with $n = 1000$ (as in (6.11) below) until convergence occurs, in which case an estimate for p is obtained. This procedure can be repeated many times by sampling repeatedly many different sets of 1000 $B_{(j)}$'s to obtain many point estimates from which interval estimates can then be constructed. This iteration process is illustrated next.

6.5.1 Illustrations of an Iterative Process

The following illustrations deal with two lognormal and two real data examples. The four cases underscore the fact that ROSF is used with a gamma tilt function while the data, at least in the lognormal cases, are not gamma distributed. Running 10,000 fusions takes about 5 minutes in R which translates to about 8 hours for 1,000,000 fusions. In what follows the p-increments at which (6.11) is evaluated are chosen mostly as $\mathcal{O}(\bar{B})$. In all cases the maxima of the approximate B-curves were larger than the true p.

6.5.1.1 Lognormal(1,1)

In this example $\boldsymbol{X}_0$ is a LN(1,1) sample where $\max(\boldsymbol{X}_0) = 25.17781$. With $T = 59.75377$ the true tail probability to be estimated is $p = 0.001$, using $n_0 = n_1 = 100$ and $\boldsymbol{h} = (x, \log x)$. The generated fusion samples $\boldsymbol{X}_1$ are from Unif(0,100), $100 > T$, and F_B was obtained from 10,000 fusions.

We first sample 1000 from 10,000 $B_{(j)}$'s to get an approximate B-curve, and then iterate between it and the smallest p such that

$$\sum_{k=0}^{j-1} \binom{1000}{k} [F_B(p)]^k [1 - F_B(p)]^{n-k} \leq 0.95 \tag{6.11}$$

evaluated at increments of $p = 0.0001$ ($\bar{B} = 0.00060$). Starting with $j = 1000$, the sequence (j, p_j) is

$$1000 \to 0.0035 \to 996 \to 0.0028 \to 985 \to 0.0022 \to 968 \to 0.0019 \to 951 \to 0.0017 \to 937 \to 0.0016 \to 929 \to 0.0015 \to 915 \to 0.0014 \to 905 \to 0.0013 \to 888 \to 0.0012 \to 871 \to 0.0012 \cdots$$

so that convergence occurs at $\hat{p} = 0.0012$ as 0.0012 gives $j = 871$ again and again. This also suggests $K = 20$ in (6.5) which gives 0.0012 as an upper bound for p. The left side of (6.11) for $j = 871$ is the step function shown in Figure 6.3.

Repeating this with a different LN(1,1) reference sample $\boldsymbol{X}_0$ such that $\max(\boldsymbol{X}_0) = 28.27287$, and fusing 10,000 times with $\boldsymbol{X}_1$ from Unif(0,80), $80 > T$, gives with p-increments of 0.0002 ($\bar{B} = 0.00031$) the (j, p_j) sequence,

$$1000 \rightarrow 0.003 \rightarrow 995 \rightarrow 0.0024 \rightarrow 991 \rightarrow 0.002 \rightarrow 986 \rightarrow 0.0018 \rightarrow 977 \rightarrow$$
$$0.0016 \rightarrow 965 \rightarrow 0.0014 \rightarrow 954 \rightarrow 0.0012 \rightarrow 941 \rightarrow 0.001 \rightarrow 923 \rightarrow 0.001 \cdots$$

so that $\hat{p} = 0.001$ is equal to the true p.

Now, convergence might be problematic when $\max(\boldsymbol{X}_0)$ is small relative to T. In that case an augmentation of the data is helpful. Thus, repeating the previous illustration with a LN(1,1) sample where $\max(\boldsymbol{X}_0) = 16.92843$, the latter is somewhat small relative to $T = 59.75377$. Indeed, $n_0 = n_1 = 100$ and p-increments of 0.0001 (although $\bar{B} = 4.661 \times 10^{-5}$, 0.00005 was not useful), gave an imprecise $\hat{p}$ for the true tail probability $p = 0.001$:

$$1000 \rightarrow 0.001 \rightarrow 999 \rightarrow 0.0008 \rightarrow 995 \rightarrow 0.0006 \rightarrow 992 \rightarrow 0.0005 \rightarrow 989 \rightarrow$$
$$0.0005 \cdots$$

Augmenting the sample with 20 additional LN(1,1) observations resulted in a larger $\max(\boldsymbol{X}_0) = 31.7835$ and $n_0 = n_1 = 120$. We have with $\boldsymbol{X}_1$ from Unif(0,100), $100 > T$, and p-incremnt=0.0001 (now $\bar{B} = 0.0003211$)

$$1000 \rightarrow 0.0038 \rightarrow 998 \rightarrow 0.0031 \rightarrow 995 \rightarrow 0.0028 \rightarrow 991 \rightarrow 0.0024 \rightarrow 987 \rightarrow$$
$$0.0022 \rightarrow 980 \rightarrow 0.0019 \rightarrow 970 \rightarrow 0.0016 \rightarrow 959 \rightarrow 0.0014 \rightarrow 951 \rightarrow 0.0013 \rightarrow$$
$$946 \rightarrow 0.0012 \rightarrow 938 \rightarrow 0.0011 \rightarrow 932 \rightarrow 0.001 \rightarrow 926 \rightarrow 0.001 \cdots$$

so that with the augmented data $\hat{p} = 0.001$ has been rendered precise.

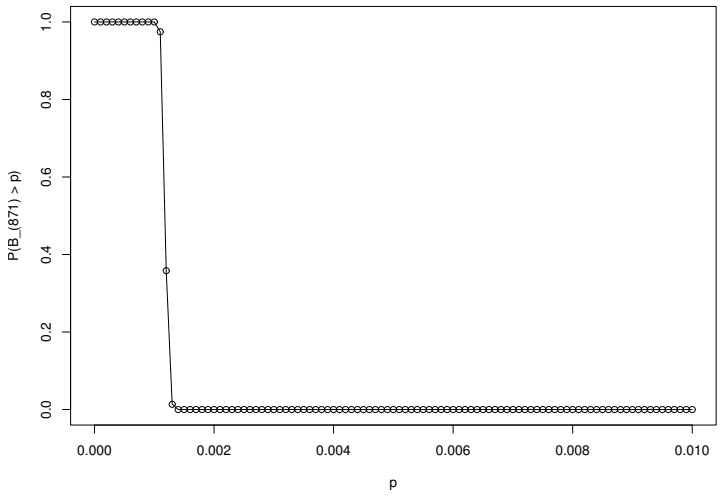

Figure 6.3: Step function (6.11) from $\boldsymbol{X}_0 \sim \text{LN}(1,1)$ fused with $\boldsymbol{X}_1 \sim \text{Unif}(0,100)$ data for $j = 871$ and containing a point corresponding to $\hat{p} = 0.0012$.

6.5.1.2 Lognormal(0,1)

Here $\boldsymbol{X}_0$ is a LN(0,1) sample where $\max(\boldsymbol{X}_0) = 5.77902$, which is small relative to $T = 21.98218$. Instead of addition of more data, we opt for more fusions.

In this example (6.11) is evaluated as a function of p using increments of 0.0002. The true tail probability is $p = 0.001$, and F_B was obtained from 1,000,000 fusions of $\boldsymbol{X}_0$ with $\boldsymbol{X}_1$ from Unif(0,40), $40 > T$. Again $n_0 = n_1 = 100$ and $\boldsymbol{h} = (x, \log x)$. In this example the largest point estimate of p from one million point estimates (obtained from (5.1) in Section 1.5) was only 0.0004186, much lower than the true p, and the reason why we use the B_j upper bounds. Again, first 1000 $B_{(j)}$'s were

sampled at random from 1,000,000 $B_{(j)}$'s to get an approximate B-curve with 1000 points $(j, B_{(j)})$. Starting with the maximum $j = 1000$ the sequence (j, p_j) along p-increments of 0.00002 ($\bar{B} = 0.000065$) is:

$$1000 \to 0.001 \to 1000 \to 0.001 \to 1000 \to 0.001 \cdots$$

so that the convergence gives the exact $p = 0.001$. This suggests $K = 2100$ in (6.5) giving an upper bound of 0.0011. Note that K is large as $\max(\boldsymbol{X}_0)$ is small relative to T.

Repeating this with 100,000 fusions, $\max(\boldsymbol{X}_0) = 7.510843$, and $\boldsymbol{X}_1$ from Unif(0,30), $30 > T$, gives the (j, p_j) sequence along p-increments of 0.0002 ($\bar{B} = 0.00014$)

$$1000 \to 0.0016 \to 998 \to 0.0014 \to 997 \to 0.0012 \to 994 \to 0.001 \to 991 \to 0.001 \cdots$$

which again converges to the exact $p = 0.001$. This suggests $K = 300$ in (6.5), giving an upper bound of 0.0011.

6.5.1.3 Mercury

Here $\boldsymbol{X}_0$ is a sample of size $n_0 = 100$ from the mercury data whose histogram is shown in Figure 6.2. Again $n_1 = 100$ and $\boldsymbol{h} = (x, \log x)$. The mercury data

consist of 8266 observations for which $T = 22.41$ gives $p = 0.001088797 \approx 0.001$. We have $\max(\boldsymbol{X}_0) = 9.09$, $\boldsymbol{X}_1 \sim \text{Unif}(0, 40)$, $40 > T$, and F_B was obtained from 1,000,000 fusions. Sampling 1000 $B_{(j)}$'s from 1,000,000 $B_{(j)}$'s, the (j, p_j) sequence along p-increments of 0.0002 ($\bar{B} = 0.00096$) is:

$$1000 \to 0.0052 \to 996 \to 0.0046 \to 991 \to 0.0042 \to 981 \to 0.0038 \to 966 \to 0.0034 \to 949 \to 0.0032 \to 942 \to 0.0030 \to 911 \to 0.0026 \to 895 \to 0.0024 \to 879 \to 0.0022 \to 851 \to 0.0020 \to 829 \to 0.0018 \to 801 \to 0.0016, \to 768 \to 0.0014 \to 732 \to 0.0014 \cdots$$

converging to $\hat{p} = 0.0014$, exactly what we get with $K = 10$ in (6.5).

Starting with a $B_{(j)}$ closer to the true $p = 0.001$ we get an upward convergence,

$$637 \to 0.001 \to 651 \to 0.001$$

.

In a different run with only 10,000 $B_{(j)}$, $\max(\boldsymbol{X}_0) = 11$, $\boldsymbol{X}_1 \sim \text{Unif}(0, 50)$, $50 > T$, and sampling 1000 $B_{(j)}$'s from 10,000 $B_{(j)}$'s, the (j, p_j) sequence along p-increments of 0.0001 (although $\bar{B} = 0.001331$) is:

$$745 \to 0.0019 \to 722 \to 0.0017 \to 695 \to 0.0015 \to 657 \to 0.0013 \to 634 \to 0.0012 \to 617 \to 0.0011 \to 606 \to 0.001 \to 589 \to 0.001 \cdots$$

Consider the higher probability $p = 0.01004113$ corresponding to $T = 9.375$, and a mercury sample $\boldsymbol{X}_0$ of size $n_0 = 100$ where $\max(\boldsymbol{X}_0) = 7.77 < T$, and $\boldsymbol{X}_1 \sim \text{Unif}(0, 20)$, $20 > T$, $n_1 = 100$. Out of 10,000 fusions with $\boldsymbol{h}(x) = (x, \log x)$, giving a different F_B than the previous one, the maximum probability estimate (out of 10,000) using (5.1) is 0.003738044, far below the true $p = 0.01004113$, and the maximum likelihood estimate based on $\boldsymbol{X}_0$ only is 0. On the other hand, sampling 1000 $B_{(j)}$'s from 10,000 $B_{(j)}$'s, the ROSF iterative (j, p_j) sequence along p-increments of 0.001 ($\bar{B} = 0.002686784$) is:

$$1000 \to 0.011 \to 1000 \to 0.011 \cdots$$

while starting from $B_{(999)}$ yields

$$999 \to 0.01 \to 996 \to 0.01 \cdots$$

so that $\hat{p} \approx p$. $K = 800$ in (6.5) gives 0.0102 as an upper bound.

6.5.1.4 Lead Intake

Here $\boldsymbol{X}_0$ is a sample of size $n_0 = 100$ from the lead intake data whose histogram is shown in Figure 6.2. Again $n_1 = 100$ and $\boldsymbol{h} = (x, \log x)$. The lead data consist of 3000 observation for which $T = 25$ gives $p = 0.001$. We have $\max(\boldsymbol{X}_0) = 11.55768$, $\boldsymbol{X}_1 \sim \text{Unif}(0, 40)$, $40 > T$, and F_B was obtained from 10,000 fusions. Again 1000 $B_{(j)}$'s were sampled from 10,000 $B_{(j)}$'s and the (j, p_j) sequence was observed along p-increments of 0.0001. In this example the iteration process starts with $j = 400$ giving p_{400} not far from the true $p = 0.001$. We have:

$$400 \to 0.0017 \to 371 \to 0.0016 \to 351 \to 0.0015 \to 327 \to 0.0014 \to 302 \to$$
$$0.0013 \to 278 \to 0.0012 \to 252 \to 0.0011 \to 229 \to 0.0011 \cdots$$

Thus, the sequence p_j converges to $\hat{p} = 0.0011$. This corresponds to $K = 2$ in (6.5).

Figure 6.4 shows the step function (6.9) for $n = 1000$ when convergence occurs at $j = 229$. Observe that $\hat{p} = 0.0011$ is the smallest p satisfying (6.11), giving a point on the cord corresponding to the pair $(0.0011, 0.3648204)$.

Now, let us see what happens in neighborhood of true p=0.001. We have:

$$201 \to 0.001 \to 203 \to 0.001 \cdots$$

and the convergence is upward. With

$$205 \to 0.001 \to 203 \to 0.001 \cdots$$

the convergence is downward. This shows that in a neighborhood of the true p the $B_{(j_k)}$ can change course to lock on the true, or approximately true, p from above or from below.

Consider the higher probability $p = 0.01$ corresponding to $T = 10$, and a lead intake sample $\boldsymbol{X}_0$ of size $n_0 = 100$ where $\max(\boldsymbol{X}_0) = 6.875607 < T$, and $\boldsymbol{X}_1 \sim \text{Unif}(0, 20)$, $20 > T$, $n_1 = 100$. Out of 10,000 fusions with $\boldsymbol{h}(x) = (x, \log x)$, giving F_B, the maximum probability estimate (out of 10,000) using (5.1) is 0.003550, far below the true $p = 0.01$, and the maximum likelihood estimate based on $\boldsymbol{X}_0$ only

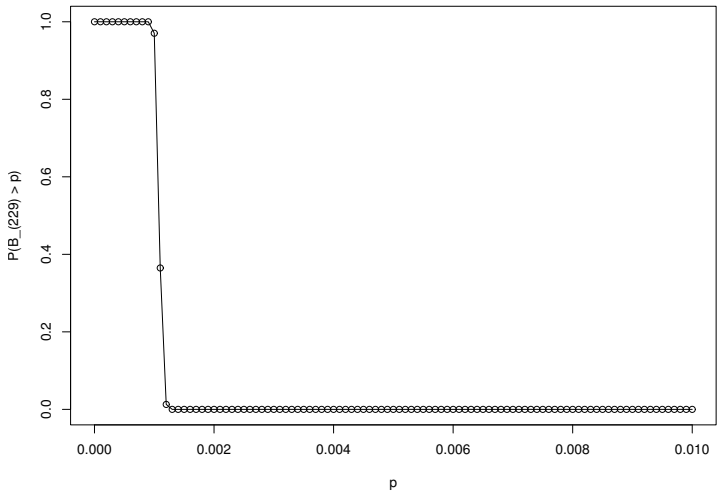

Figure 6.4: Step function (6.11) from lead intake $\boldsymbol{X}_0$ fused with $\boldsymbol{X}_1 \sim \text{Unif}(0, 40)$ data for $j = 229$ and containing a point corresponding to $\hat{p} = 0.0011$ whose ordinate is $0.3648204 < 0.95$.

is 0. On the other hand, sampling 1000 $B_{(j)}$'s from 10,000 $B_{(j)}$'s, the IM iterative (j, p_j) sequence along p-increments of 0.001 ($\bar{B} = 0.003516579$) is:

$$1000 \to 0.01 \to 999 \to 0.009 \to 998 \to 0.009 \cdots$$

so that $\hat{p} = 0.009$. We note that $K = 5000$ in (6.5) gives 0.01.

6.5.2 Explaining the Convergence

Clearly, the p_j sequence need not converge to a point in a neighborhood of the true p. However, as we have seen from the previous illustrations, iterating between the two monotone relationships, the B-curve and (6.11) along certain p-increments, we do get in many cases p estimates in neighborhoods of the true p. This can be explained as follows.

From the previous illustrations we observe that when starting with a sufficiently large j_1 we have a monotone decreasing sequence,

$$B_{j_1} > B_{j_2} > B_{j_3} \cdots ,$$

and suppose that, for some j, $B_{(j)}$ lands in a neighborhood of the true p. As $n \to \infty$, the $B_{(j)}$'s become ever more dense so that the absolute difference $\left|B_{(j\pm k)} - B_{(j)}\right|$ becomes arbitrarily small for $B_{(j\pm k)}$ in that neighborhood. *Therefore, the smallest p's in that neighborhood, along certain p-increments, which satisfy (e.g. as in (6.11))*

$$P(B_{(j)} > p) \leq 0.95 \tag{6.12}$$

are equal or nearly equal for entire stretches of adjacent $B_{(j)}$'s, thus increasing the probability that two successive p_j's in the iteration process are equal, in which case

convergence occurs in a neighborhood of the true p.

This can be illustrated with $\boldsymbol{X}_0 \sim LN(0,1)$, $\max(\boldsymbol{X}_0) = 9.274283$, $\boldsymbol{X}_1 \sim$ Unif$(0, 30)$, $30 > T = 21.98218$, $p = 0.001$, $n_0 = n_1 = 100$, p-increments of 0.0001 ($\bar{B} = 0.000414989$), and 30,000 B_j. Sampling 1000 B_j, successive $B_{(j)}$ in a neighborhood of $p = 0.001$ give:

For $957 \leq j \leq 950$, the smallest p_j which satisfies (6.11) is 0.0011.

For $949 \leq j \leq 938$, the smallest p_j which satisfies (6.11) is 0.0010.

For $937 \leq j \leq 924$, the smallest p_j which satisfies (6.11) is 0.0009.

Hence, over the stretch $j = 957$ to $j = 924$ there are 34 consecutive p_j's which are markedly close to the true $p = 0.001$. As a result, in this stretch, starting with $j = 957$ along increments of 0.0001 the next j in the iteration process is $j = 950$ and we have two equal successive p_j,

$$957 \to 0.0011 \to 950 \to 0.0011,$$

whereas starting with $j = 949$ the next j's in the iteration process are $j = 937$ and $j = 926$, and again there are two equal successive p_j,

$$949 \to 0.001 \to 937 \to 0.0009 \to 926 \to 0.0009.$$

Thus, entering a neighborhood of the true p, the iteration method (IM) produces further p's which, as n increases, tend to stay in that neighborhood leading to convergence. We have seen this tendency throughout the previous illustrations, and we see more of it from the tables in the next section.

6.6 Comparison: ROFS vs POT

Against the background provided in the previous sections, we compare two very different ways to obtain interval estimates for small tail probabilities: POT based on extreme value theory, and an iterative process based on repeated fusion of a given reference sample with external computer-generated uniformly distributed samples. The comparison is based on confidence interval coverage, width, and on the mean absolute error (MAE) which measures the discrepancy between $\hat{p}$ and the true tail probability p. In Tables 6.1 to 6.12, p is relatively small, $p = 0.001$, whereas in the last three tables 6.13, 6.14, 6.15, p is smaller, $p = 0.0001$.

Throughout the comparison the sample sizes are $n_0 = n_1 = 100$ or $n_0 = n_1 = 200$, and $\boldsymbol{h}(x) = (x, \log x)$. Thus, in the present comparison the reference $\boldsymbol{X}_0$ and the fusion samples $\boldsymbol{X}_1$ have size $n_0 = 100$ or $n_0 = 200$.

To save computation time, in each case of the iteration process F_B was obtained from 1000 fusions, and the starting j is such that $B_{(j)}$ is approximately equal to the 3rd quartile of the observed 1000 B's.

Remark: Starting at the 3rd quartile is computationally sensible as the corresponding $B_{(j)}$ is usually in a neighborhood above p. In most cases subsequent $B_{(j)}$ do enter a neighborhood of p and convergence occurs, as explained earlier. Starting too low might lead to convergence to a point lower than the true p.

The following tables are the result of 500 runs. In each run the iteration method

(IM) was repeated 500 times.

From the mean residual life (MRL) plots we obtained the thresholds u needed for the POT method. In all cases reported in the tables, the MRL plots suggest the use of the largest 20% of the reference data $\boldsymbol{X}_0$ for fitting the generalized Pareto (GP) distribution. We have noticed a deterioration in the POT results when using 30%, 15% or 10% of $\boldsymbol{X}_0$. The simulation details are given in Appendix A.

An interesting picture emerges from Tables 6.1 to 6.15. For moderately large sample sizes of $n_0 = 100$ and $n_0 = 200$, regardless of the tail type, as N, the number of $\hat{p}$'s used in forming the CI for the true p, grows the iteration process gives reliable and relatively narrow confidence intervals, whereas the POT gives unacceptable coverage and in many cases wider CI's as well. The POT coverage increases significantly going from $n_0 = 100$ to $n_0 = 200$, however, it seems that for the method to "fire up" larger samples are needed. Regarding ROSF, the choice of $\mathbf{N = 50}$ (N is defined above and in Appendix A) seems prudent across all cases, and with $n_0 = 200$ shorter CI's achieve coverage similar to that from $n_0 = 100$. In all cases the MAE from the iteration process is much smaller than that obtained from POT.

6.6.1 Comparison Tables

The following tables compare ROSF and POT for $p = 0.001$ and $p = 0.0001$.

Table 6.1: $X_0 \sim \mathbf{t}_{(1)} : p = 1 - G(T) = 0.001, T = 631.8645, X_1 \sim$ Unif(0,800), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.0001.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	63.2%	0.00372	0.00149	72.1%	0.00292	0.00122
ROSF & IM	5	47.2%	0.00098	0.00061	54.1%	0.00079	0.00051
	10	57.2%	0.00107	-	68.5%	0.00093	-
	25	74.3%	0.00148	-	87.2%	0.00125	-
	50	98.2%	0.00213	-	100%	0.00193	-
	100	100%	0.00264	-	100%	0.00241	-
	300	100%	0.00321	-	100%	0.00303	-

Table 6.2: $X_0 \sim \mathbf{Weibull}(1, 2) : p = 1 - G(T) = 0.001, T = 13.81551, X_1 \sim$ Unif(0,16), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.00005.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	82.7%	0.00431	0.00131	87.8%	0.00333	0.00083
ROSF & IM	5	43.2%	0.00040	0.00068	52.4%	0.00042	0.00051
	10	65.2%	0.00083	-	72.7%	0.00091	-
	25	84.2%	0.00159	-	85.6%	0.00154	-
	50	92.5%	0.00287	-	92.8%	0.00231	-
	100	100%	0.00381	-	100%	0.00321	-
	300	100%	0.00506	-	100%	0.00402	-

Table 6.3: $X_0 \sim \mathbf{Pareto}(1, 4) : p = 1-G(T) = 0.001, T = 5.623413, X_1 \sim$ Unif(1,8), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.0001.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	81.8%	0.00419	0.00121	84.5%	0.00337	0.00070
ROSF & IM	5	59.1%	0.00068	0.00052	62.4%	0.00066	0.00041
	10	66.7%	0.00093	-	74.8%	0.00091	-
	25	84.1%	0.00154	-	86.1%	0.00148	-
	50	96.2%	0.00232	-	97.8%	0.00231	-
	100	100%	0.00272	-	100%	0.00269	-
	300	100%	0.00397	-	100%	0.00377	-

Table 6.4: $X_0 \sim \mathbf{Gamma}(3, 1) : p = 1 - G(T) = 0.001, T = 11.22887, X_1 \sim$ Unif(0,20), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.00005.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	77.3%	0.00410	0.00096	86.1%	0.00321	0.00081
ROSF & IM	5	41.2%	0.00057	0.00054	47.1%	0.00056	0.00043
	10	49.6%	0.00093	-	56.6%	0.00092	-
	25	73.2%	0.00137	-	82.8%	0.00129	-
	50	93.4%	0.00188	-	94.5%	0.00175	-
	100	100%	0.00256	-	100%	0.00248	-
	300	100%	0.00338	-	100%	0.00315	-

Table 6.5: $X_0 \sim \mathbf{F}(2,12) : p = 1 - G(T) = 0.001, T = 12.97367, X_1 \sim$ Unif(0,16), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.00005.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	83.1%	0.00372	0.00111	87.0%	0.00292	0.00082
ROSF & IM	5	43.1%	0.00066	0.00051	46.1%	0.00058	0.00031
	10	54.2%	0.00094	-	58.1%	0.00088	-
	25	78.5%	0.00136	-	83.5%	0.00131	-
	50	96.1%	0.00217	-	98.6%	0.00189	-
	100	100%	0.00289	-	100%	0.00277	-
	300	100%	0.00344	-	100%	0.00323	-

Table 6.6: $X_0 \sim \mathbf{IG}(2,40) : p = 1 - G(T) = 0.001, T = 3.835791, X_1 \sim$ Unif(0,8), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.00005.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	69.6%	0.00324	0.00123	82.3%	0.00316	0.00092
ROSF & IM	5	59.4%	0.00098	0.00047	85.6%	0.00091	0.00041
	10	76.9%	0.00147	-	96.3%	0.00133	-
	25	89.9%	0.00255	-	100%	0.00147	-
	50	100%	0.00289	-	100%	0.00206	-
	100	100%	0.00332	-	100%	0.00313	-
	300	100%	0.00401	-	100%	0.00371	-

Table 6.7: $X_0 \sim \mathbf{IG}(4,5) : p = 1 - G(T) = 0.001, T = 28.95409, X_1 \sim$ Unif(0,35), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.00005.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	84.3%	0.00412	0.00123	88.9%	0.00339	0.00103
ROSF & IM	5	76.3%	0.00106	0.00052	80.4%	0.00087	0.00041
	10	89.2%	0.00148	-	87.1%	0.00127	-
	25	97.5%	0.00217	-	98.9%	0.00172	-
	50	100%	0.00265	-	100%	0.00225	-
	100	100%	0.00345	-	100%	0.00259	-
	300	100%	0.00372	-	100%	0.00291	-

Table 6.8: $X_0 \sim \mathbf{LN}(0,1) : p = 1 - G(T) = 0.001, T = 21.98218, X_1 \sim$ Unif(1,60), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.00005.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	81.5%	0.00451	0.00111	85.2%	0.00392	0.00103
ROSF & IM	5	81.5%	0.00121	0.00047	83.6%	0.00108	0.00039
	10	88.7%	0.00169	-	90.4%	0.00141	-
	25	95.3%	0.00191	-	98.1%	0.00173	-
	50	100%	0.00234	-	100%	0.00199	-
	100	100%	0.00267	-	100%	0.00244	-
	300	100%	0.00301	-	100%	0.00283	-

Table 6.9: $X_0 \sim \mathbf{LN}(1,1) : p = 1 - G(T) = 0.001, T = 59.75377, X_1 \sim$ Unif(1,140), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.0001.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	81.4%	0.00435	0.00117	86.8%	0.00399	0.00099
	5	43.7%	0.00078	0.00069	53.2%	0.00071	0.00052
	10	56.9%	0.00109	-	68.1%	0.00099	-
	25	79.6%	0.00143	-	89.7%	0.00121	-
ROSF & IM	**50**	89.1%	0.00187	-	100%	0.00164	-
	100	100%	0.00199	-	100%	0.00192	-
	300	100%	0.00243	-	100%	0.00234	-

Table 6.10: $X_0 \sim$ Mercury : $p = 1 - G(T) = 0.001, T = 22.41, X_1 \sim$ Unif(0,50), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.0001.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	85.3%	0.00455	0.00130	88.6%	0.00398	0.00122
	5	54.5%	0.00073	0.00048	49.9%	0.00063	0.00045
	10	66.7%	0.00095	-	76.7%	0.00096	-
	25	84.9%	0.00157	-	96.7%	0.00145	-
ROSF & IM	**50**	97.5%	0.00215	-	100%	0.00197	-
	100	100%	0.00259	-	100%	0.00238	-
	300	100%	0.00337	-	100%	0.00313	-

Table 6.11: $X_0 \sim$ Lead Intake : $p = 1 - G(T) = 0.001, T = 25, X_1 \sim$ Unif(0,30), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.0001.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	84.7%	0.00555	0.00142	87.7%	0.00536	0.00125
	5	51.1%	0.00095	0.00066	49.6%	0.00088	0.00058
	10	69.3%	0.00151	-	78.1%	0.00153	-
	25	88.4%	0.00189	-	93.7%	0.00179	-
ROSF & IM	**50**	100%	0.00247	-	100%	0.00229	-
	100	100%	0.00289	-	100%	0.00268	-
	300	100%	0.00346	-	100%	0.00317	-

Table 6.12: $X_0 \sim$ URX3TB : $p = 1 - G(T) = 0.001, T = 9.50, X_1 \sim$ Unif(0,12), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.0001. Data source for URX3TB - 2,4,6-trichlorophenol (ug/L): `https://wwwn.cdc.gov/nchs/nhanes`

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	81.1%	0.00433	0.00143	87.1%	0.00376	0.00123
	5	38.9%	0.00078	0.00055	42.6%	0.00071	0.00044
	10	54.3%	0.00094	-	61.8%	0.00092	-
	25	72.1%	0.00131	-	81.7%	0.00125	-
ROSF & IM	**50**	89.1%	0.00179	-	96.9%	0.00177	-
	100	100%	0.00241	-	100%	0.00235	-
	300	100%	0.00264	-	100%	0.00259	-

Table 6.13: $X_0 \sim \mathbf{F}(2, 12) : p = 1 - G(T) = 0.0001, T = 21.84953, X_1 \sim$ Unif(0,25), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.00001.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	71.4%	0.00062	0.00052	81.6%	0.00053	0.000045
	5	45.2%	0.00021	0.00022	49.1%	0.00017	0.000019
	10	67.2%	0.00033	-	77.1%	0.00026	-
	25	88.5%	0.00045	-	89.3%	0.00037	-
ROSF & IM	**50**	95.2%	0.00059	-	96.3%	0.00052	-
	100	100%	0.00082	-	100%	0.00069	-
	300	100%	0.00105	-	100%	0.00087	-

Table 6.14: $X_0 \sim \mathbf{LN}(0, 1) : p = 1 - G(T) = 0.0001, T = 41.22383, X_1 \sim$ Unif(1,60), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.00001.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	72.1%	0.00064	0.00045	82.6%	0.00047	0.000039
	5	55.2%	0.00021	0.00021	69.1%	0.00017	0.000017
	10	77.2%	0.00033	-	89.1%	0.00020	-
	25	98.5%	0.00041	-	99.3%	0.00034	-
ROSF & IM	**50**	100%	0.00066	-	100%	0.00057	-
	100	100%	0.00083	-	100%	0.00079	-
	300	100%	0.00113	-	100%	0.00094	-

Table 6.15: $X_0 \sim$ Mercury : $p = 1 - G(T) = 0.0001, T = 39.60, X_1 \sim$ Unif(0,80), $n_0 = n_1, h(x) = (x, \log x)$. p-increment 0.00001.

		$n_0 = 100$			$n_0 = 200$		
Method	N	Coverage	CI Length	MAE	Coverage	CI Length	MAE
POT	-	62.4%	0.00059	0.00049	73.4%	0.00051	0.000042
	5	53.1%	0.00019	0.00023	64.2%	0.00016	0.000019
	10	71.8%	0.00025	-	79.8%	0.00021	-
	25	88.3%	0.00037	-	91.5%	0.00033	-
ROSF & IM	**50**	95.2%	0.00056	-	100%	0.00054	-
	100	100%	0.00083	-	100%	0.00079	-
	300	100%	0.00113	-	100%	0.00094	-

6.7 Discussion

The numerous number of fusions of a given reference sample with computer generated samples gives rise to different observables including the upper bounds for a tail probability p that were used in the Chapter. The upper bounds, obtained from the combined real and artificial data, were mostly much larger than p, some were less than p, but some among the multitude of upper bounds essentially coincided with p and they were identified to a reasonable degree of approximation using an iteration procedure.

We have illustrated that repeated fusion of a sample with generated uniform random data allowed us to gain information about the tail behavior beyond the threshold using the notion of B-curves coupled with a well known formula from order statistics.

The following example summarizes our ideas. Consider the B-curve in Figure 6.5. It was obtained from a LN(1,1) sample of size $n_0 = 200$, fused 10,000 times with independent computer generated Unif(0,100) samples each of size $n_1 = 200$. The curve contains a point whose ordinate is the tail probability $p = 0.001$ which we wish to estimate. From the curve we see immediately that $B_{(1)} < p < B_{(10,000)}$ or, approximately, $0 < p < 0.003$. That is, ROSF gives a useful and fast interval estimate for p. In most cases, the iteration method (IM) refines this estimate. To see this in the present case, starting with $j = 1,000$, the IM convergence results from ten different $B_{(j)}$-samples of size 1,000 obtained from 10,000 $B_{(j)}$'s were 0.002, 0.0009, 0.001, 0.001, 0.0012, 0.0007, 0.0009, 0.001, 0.002, 0.001 with an average of

0.00117, not far from $p = 0.001$, and a standard deviation of 0.00045. This example points to the fact that IM can be repeated many times with different $B_{(j)}$-samples to produce tail probability estimates and their precision.

In this dissertation, we have discussed three different types of data fusion based on the Density Ratio Model. We have seen by integrating multiple data sources, we can produce more accurate, and useful information than that provided by individual data source. In the small area problem, some natural extensions could consider in the future research. They include improving the model structure, more flexibility of the linear component in the model assumption, and fusing with different type of error distribution. The ideas presented in the tail probability estimation can also be extended in a number of ways. For example, using "fake" data from distributions other than uniform, and using different fusion mechanisms other than the semiparametric method. Estimating K in (6.5) is another possible extension.

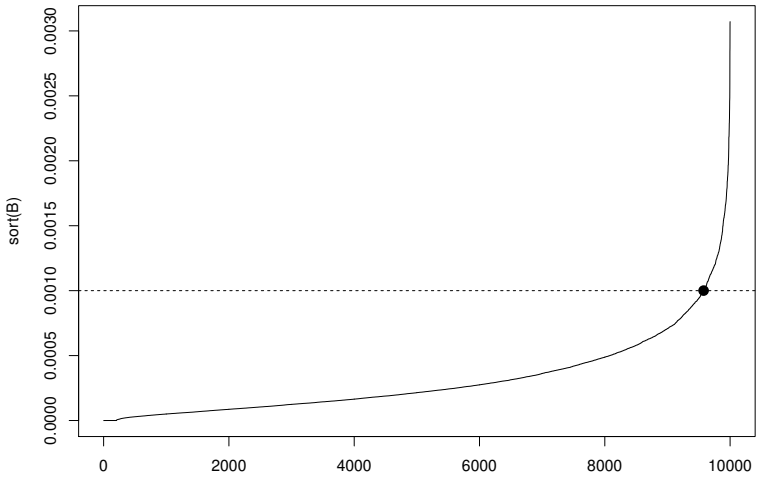

Figure 6.5: B-Curve containing a point corresponding to $p = 0.001$, obtained from a reference LN(1,1) sample fused 10,000 times with independent Unif(0,100) samples. $\boldsymbol{h}(x) = (x, \log x), \max(\boldsymbol{X}_0) = 25.46234, T = 59.7538, n_0 = n_1 = 200$.

Appendix A: Simulation Description

The following steps were followed. There were 500 runs. In each run the iteration method (IM) was repeated 500 times.

First, a reference $\boldsymbol{X}_0$ was obtained.

POT:

The POT procedure was applied to get both an estimate $\hat{p}$ and a confidence interval (CI). The MRL plots suggest the use of the largest 20% of the reference data $\boldsymbol{X}_0$ for fitting the generalized Pareto (GP) distribution.

ROSF/IM:

$\boldsymbol{X}_0$ was fused with $\boldsymbol{X}_1$ 1000 times (ROSF) to get F_B and then $\hat{p}$ (IM).

$\boldsymbol{X}_0$ was fused again with different $\boldsymbol{X}_1$ 1000 times to get F_B and $\hat{p}$.

This was repeated 500 times.

The iterative method thus gave 500 $\hat{p}$'s. We then chose at random N $\hat{p}$'s from 500 $\hat{p}$'s to construct a CI for the true p as $(\min(\hat{p}), \max(\hat{p}))$.

This is run 1.

The above steps were repeated, for both POT and ROSF/IM each time with a different $\boldsymbol{X}_0$, 500 times (runs) to obtain coverage and average CI length. In the tables, CI length is an average length from 500 intervals.

Since there are 500 runs, POT gave 500 $\hat{p}$'s. Regarding IM, a single $\hat{p}$ was chosen at random (out of 500 $\hat{p}$'s) from each of the 500 runs. The mean absolute error (MAE) was obtained in both cases from the mean of 500 absolute differences $\sum(|\hat{p}_i - p|)/500$, where $p = 0.001$ or $p = 0.0001$. In the iterative method, in each table the MAE is reported once on the line corresponding to $N = 5$.

Bibliography

[1] Battese, G. E., Harter, R. M. and Fuller, W. A. (1988), *An error-components model for prediction of county crop areas using survey and satellite data*, Journal of the American Statistical Association, 80, 28-36.

[2] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J., (2004). *Statistics of Extremes Theory and Application*, Wiley.

[3] Balkema, A., and de Haan, L. (1974). *Residual life time at great age*, Annals of Probability, 2, 792–804.

[4] Cristalli, C., Paone, N. and Rodriguez, R.M. (2006). *Mechanical Fault Detection of Electric Motors by Laser Vibrometer and Accelerometer Measurements*, Mechanical Systems and Signal Processing, **20**, 1350-1361

[5] Concettoni, E., Cristalli, C., and Serani, S. (2012). *Mechanical and electrical quality control tests for small dc motors in production line*, IECON 2012-38th Annual Conference of IEEE Industrial Electronics Society, 1883-1887.

[6] Chen, J. and Liu, Y. (2015). *Small Area Estimation under Density Ratio Model*, Research paper.

[7] Coles, S., (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer.

[8] David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*, Wiley Series in Probability and Statistics.

[9] de Haan, L. and Ferreira, A., (2006). *Extreme Value Theory An Introduction*, Springer.

[10] Ferreira, A. and De Haan, L. (2015). *On the block maxima method in extreme value theory: PWM estimators*, The Annals of Statistics, 43: 276298.

[11] Fokianos, K., Kedem, B., Qin, J. , Short, D. (2001). *A semiparametric approach to the one-way layout*, Technometrics, 43, 56-65.

[12] Fokianos, K. (2004). *Merging information for semiparametric density estimation*, JRSS, B, 66, 941-958.

[13] Fokianos, K. and Qin J. (2008). *A Note on Monte Carlo Maximization by the Density Ratio Model*, Journal of Statistical Theory and Practice; 2: 355-367.

[14] Fisher, R.A. and Tippett, L.H.C. (1928). *Limiting forms of the frequency distribution of the largest or smallest member of a sample*, Proceedings of the Cambridge Philosophical Society, 24: 180-190.

[15] Fithian, W. and Wager, S. (2015). *Semiparametric exponential families for heavy-tailed data*, Biometrika, 102: 486-493.

[16] Gnedenko, B. V. (1948). *On a local limit theorem of the theory of probability*, Uspekhi Mat. Nauk, 3:3(25), 187–194

[17] Goyal, D. and Pabla, B.S. (2016). *The vibration monitoring methods and signal processing techniques for structural health monitoring: A review*, Archives of Computational Methods in Engineering, **23**, 585-594.

[18] Gilbert, P.B., Lele, S.R. and Vardi, Y. (1999). *Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials*, Biometrika, 86, 27-43.

[19] Kedem, B., Pan, L., Smith, P., Wang, C., (2018). *Repeated out of Sample Fusion in the Estimation of Small Tail Probabilities*, arXiv:1803.10766

[20] Kedem, B., Lu, G., and Williams, P. D. (2008). *Forecasting Mortality Rates Via Density Ratio Modeling*, Canadian Journal of Statistics, 36, 193-206.

[21] Kedem, B., De Oliveira, V., and Sverchkov, M. (2017). *Statistical Data Fusion*, World Scientific, Singapore.

[22] Kedem, B., Wolff, D., and Fokianos, K. (2004). *Statistical comparison of algorithms*, IEEE Tr. on Instrumentation and Measurement, 2004, 53, 770-776.

[23] Kedem, B., Pan, L., Zhou, W., and Coelho, C.A.(2016). *Interval estimation of small tail probabilities – application in food safety*, Statistics in Medicine; 35: 3229-3240.

[24] Katzoff, M., Zhou, W., Khan, D., Lu, G., Kedem, B. (2014). *Out of Sample Fusion in Risk Prediction*, Journal of Statistical Theory and Practice, 8, 3, 444-459.

[25] Koenker, Roger (2005). *Quantile Regression*, Cambridge University Press.

[26] Lu, G. (2007). *Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models and its Application to Mortality Forecasting*. Ph.D. Dissertation, University of Maryland, College Park.

[27] Leadbetter, M., Lindgren, G., Rootzen, H., (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer.

[28] Owen, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton.

[29] Pan, L. (2016). *Semiparametric Methods in the Estimation of Tail Probabilities and Extreme Quantiles*. Ph.D. Dissertation, University of Maryland, College Park.

[30] Pickands, J. (1975). *Statistical inference using extreme order statistics*, Annals of Statistics, 3, 119–131.

[31] Qin, J., Zhang, B. (1997). *A Goodness of Fit Test for Logistic Regression Models Based on Case-control Data*, Biometrika, 84, 609-618.

[32] Qin, J., Lawless, J. (1994). *Empirical Likelihood and General Estimating Equations*, The Annals of Statistics, 22, No. 1 300.

[33] Resnick, S.,(1987). *Extreme Values, Point Processes and Regular Variation*, Springer.

[34] Voulgaraki, A., Kedem, B., and Graubard, B. I. (2012). *Semiparametric Regression in Testicular Germ Cell Data*, Annals of Applied Statistics, 6, 3 1185-1208.

[35] Vardi, Y. (1985). *Empirical Distributions in Selection Bias Models*, The Annals of Statistics 13, No. 1, 178.

[36] Vardi, Y. (1982). *Nonparametric estimation in the presence of length bias*, Annals of Statistics, Vol. 10, 616-20.

[37] Yu, L. (2017). *Two Goodness-of-Fit Tests for the Density Ratio Model*. Ph.D. Dissertation, University of Maryland, College Park.

[38] Zhang, B.(2000). *A goodness of fit test for multiplicative-intercept risk models based on case-control data*, Statistica Sinica 10, 839-865.

[39] Zhou, W. (2013). *Out of Sample Fusion*. Ph.D. Dissertation, University of Maryland, College Park.